

A Faithful Explainable Adaptive Assessment Framework Using Coupled Knowledge-Behaviour Belief Modelling for Android-Based Learning Applications

**Naga Gopi Chappidi¹, Kumar N S L Prakash Gandham²,
Hanish Venkat Sundarapalli³**

¹Koneru Lakshmaiah Education Foundation (KLEF), Andhra Pradesh, India

^{2,3}Velagapudi Ramakrishna Siddhartha Engineering College (VRSEC), Vijayawada, Andhra Pradesh, India

Abstract:

Adaptive assessment systems increasingly combine prediction, learner-facing explanations, and automated instructional actions, but these components are often produced by separate mechanisms. This creates a risk that the explanation shown to a learner describes one reason for difficulty while the adaptive policy responds to another. This paper introduces the Faithful Explainable Adaptive Assessment Framework (FEAAF), centered on Explanation-Action Faithfulness: the instructional action selected by the system should remain consistent with the explanation presented to the learner. FEAAF uses a Coupled Knowledge-Behaviour Belief Model (CKBM) that tracks both skill mastery and behavioural signals such as time pressure and disengagement, because either signal alone may be insufficient to distinguish different sources of underperformance. When underperformance is detected, the framework considers a small set of diagnostic categories—knowledge gap, time pressure, and disengagement—and restricts ECAP (Explanation-Consistent Adaptive Policy) to actions compatible with the selected explanation. This restriction provides a structural guarantee: the selected action cannot contradict the explanation by construction. Because the framework is intended for Android and other on-device settings, mastery and behavioural estimates are maintained with a lightweight update strategy rather than a full joint representation across all possible skill combinations. The paper presents FEAAF as a conceptual and formal framework in relation to knowledge tracing, item-response testing, intelligent tutoring systems, and explainable or attribution-based educational data mining. Within the literature reviewed here, no comparable approach was identified that explicitly enforces explanation-action consistency at the action-selection stage. No implementation, dataset, or empirical evaluation is reported; the contribution is the framework and the faithfulness property it is designed to provide.

Keywords: Explainable AI; Adaptive Assessment; Learner Modelling; Diagnostic Reasoning; Android/On-Device Learning; Educational Data Mining; Intelligent Tutoring Systems.

1. Introduction

Mobile assessment applications increasingly combine adaptive behaviour with explanations. In many existing designs, however, explainability is added after a prediction has been produced, while a separate policy determines the next instructional action. Such an architecture does not require the explanation and the action to agree. For example, a system could tell a learner that a conceptual knowledge gap caused an incorrect response while its adaptive mechanism changes the time limit, which would instead correspond to a time-pressure explanation. This reflects a broader concern in explainable AI that an explanation may not faithfully represent what actually drives system behaviour [1]. FEAAF addresses this issue specifically in adaptive educational systems by asking whether the explanation presented to the learner is structurally aligned with the action selected next. This motivates the research question addressed in this paper: how can an adaptive educational system ensure that its instructional action remains structurally consistent with the explanation presented to the learner under practical on-device deployment constraints?

FEAAF is intended for Android and other on-device deployment settings in which learner-state updates are performed locally rather than by a remote server. This deployment setting motivates a practical design decision: instead of maintaining one large joint representation of every skill, the framework updates mastery and behavioural estimates in a lightweight, per-skill manner. This is intended to keep the proposed design feasible for consumer mobile hardware. Android is therefore treated as a constraint shaping the framework rather than as evidence of an implemented application; this paper reports no Android implementation, benchmark, latency measurement, or performance evaluation.

1.1 Contributions of This Work

C1. A Coupled Knowledge-Behaviour Belief Model (CKBM) that represents each learner through both a mastery estimate and a behavioural estimate, allowing the framework to consider information that either signal alone may not explain.

C2. A lightweight diagnostic step that identifies the most likely source of a learner's difficulty from three interpretable categories: knowledge gap, time pressure, and disengagement.

C3. The Explanation-Consistent Adaptive Policy (ECAP), which limits adaptive actions to those compatible with the diagnosed explanation, providing a structural guarantee that the selected action does not contradict the explanation shown to the learner.

C4. A design rationale for keeping learner-state updates lightweight, motivated specifically by the eventual goal of on-device deployment on Android hardware.

2. Related Work

Bayesian Knowledge Tracing (BKT) represents a learner's mastery of a skill as a hidden Markov process updated from a sequence of correct and incorrect responses [2]. It provides an interpretable mastery estimate, but does not explicitly represent behavioural signals or distinguish among different contextual reasons for an incorrect response. Deep Knowledge Tracing (DKT) uses a recurrent neural network and can improve predictive performance, but its internal representation is less directly interpretable, meaning that an explanation may need to be produced separately and is not automatically guaranteed to reflect the model's behaviour [3]. Item Response Theory (IRT)-based adaptive testing estimates an ability parameter and selects informative items, but likewise does not explicitly represent behavioural signals or provide a mechanism for distinguishing competing explanations for a response.

In educational systems, explainability is often implemented as an additional layer around an existing predictive model. Attribution-based educational data mining, for example, can associate a trained model with SHAP- or LIME-style feature attributions [4]. Because such explanations are generally produced after model training, they describe influential features but do not by themselves impose a constraint on the adaptive action that follows. Recent educational explainable-AI research has explored richer ways of presenting explanations, yet the explanation layer can still remain separate from the mechanism that chooses the next instructional action.

Several approaches in intelligent tutoring systems are also relevant. Cognitive Tutors trace learner actions against rule-based representations of expert performance, so a triggered rule can provide an interpretable diagnostic signal [10]. However, this does not necessarily provide a probabilistic comparison among competing explanations such as a knowledge misconception and time pressure. Knowledge-graph tutoring systems can propagate mastery information across related skills using graph-based models [11], but their learned representations may still require a separate explanation mechanism for learner-facing interpretation. Rule-based intelligent tutoring systems use inspectable if-then rules to select tasks or hints [12], yet transparency of the rule set does not by itself guarantee that the reason communicated to the learner is the same rule that determined the action.

Educational data mining and learning analytics have produced a substantial body of work on predictive and explainable performance modelling. Prior studies have examined teaching-process indicators in relation to learning outcomes [5], investigated instructional effects on performance [6], and reviewed broader applications of artificial intelligence in assessment [7]. More recent work includes Shapley-value-based interpretable performance prediction [4], systematic reviews of machine-learning approaches to student performance prediction [8], instructor-oriented evaluation models [9], and interpretability-focused regression for grade prediction [14]. Across these approaches, attribution-based explanations remain common, while cognitive-diagnostic modelling has also shown that student difficulties can be represented using a small number of interpretable categories [15]. FEAAF adopts the general idea of interpretable diagnostic categories while defining its own category set.

Taken together, the reviewed literature provides established approaches for prediction, explanation, and adaptive decision-making, but the reviewed works do not explicitly make explanation-action consistency a constraint on the action-selection stage. This literature-based observation, which is not claimed to be an exhaustive systematic survey, motivates FEAAF's central idea: treating consistency between what the system explains and what it subsequently does as an explicit, checkable design property.

3. Proposed Framework

3.1 Framework Overview

FEAAF operates as a continuous interaction loop shown in Figure 1. Each learner interaction updates a shared learner state containing mastery and behavioural estimates (Sec. 3.2). When underperformance is detected, the framework identifies the most likely explanation from a limited diagnostic space (Sec. 3.3), then restricts the next action to the actions permitted by that explanation (Sec. 3.4). This construction provides the central explanation-action faithfulness property described in Sec. 3.7. Figure 1 contrasts this approach with a conventional prediction-then-explain-then-adapt pipeline in which explanation and action are not explicitly coupled.

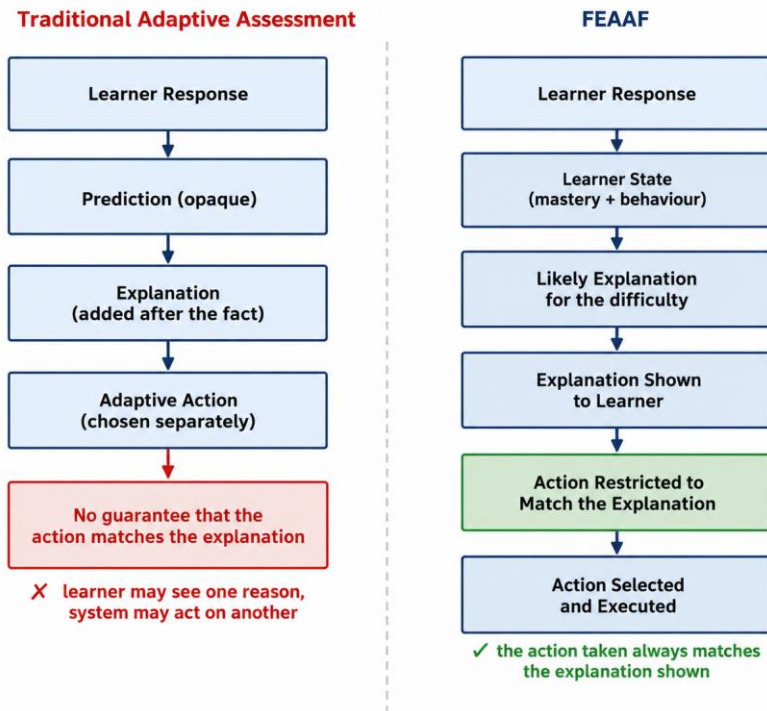


Figure 1. Traditional adaptive assessment pipelines (left) produce a prediction, add an explanation, and choose an action as separate steps, so agreement between explanation and action is not structurally guaranteed. FEAAF (right) first identifies the likely explanation and then restricts the available actions to those consistent with that explanation.

3.2 Learner State: Combining Mastery and Behaviour

Mastery information alone cannot explain every form of underperformance. A learner may have adequate knowledge of a topic but respond slowly or inconsistently because of time pressure, frustration, or disengagement, which are conceptually different from underlying competence. The control-value theory of achievement emotions provides support for distinguishing achievement-related affect and behaviour from competence itself [13]. Conversely, behavioural

observations alone cannot determine whether an unusual response pattern reflects a knowledge-related difficulty or a contextual or behavioural factor such as time pressure or disengagement. FEAAF therefore represents both dimensions in a shared learner state:

$$S_t = (M_t, B_t)$$

Here, M_t denotes a mastery estimate with one value maintained for each relevant skill, while B_t contains a small set of behavioural indicators such as time-pressure sensitivity, persistence, and engagement. Both are estimated from the same stream of learner interactions. Response correctness contributes evidence for the mastery update and, together with response timing, contributes to the behavioural update. Thus the two estimates are related through shared evidence and their joint use in diagnosis, but the mastery estimate is not directly conditioned on the behavioural estimate. This is the sense in which the representation is coupled; it does not imply direct bidirectional conditioning.

3.3 Identifying the Likely Explanation

When a learner underperforms—for example, by answering incorrectly or showing a decline on a skill—FEAAF compares the available evidence with three candidate explanations. A knowledge gap is supported when mastery of the relevant skill is low. Time-pressure-related difficulty is supported when mastery appears adequate but response timing is unusually long relative to the learner's normal pace. Disengagement is supported when engagement declines following repeated errors. The framework then selects the most likely explanation for that interaction. Because the diagnosis is derived directly from the shared learner state already maintained by FEAAF, the learner-facing explanation reflects the same diagnostic reasoning used to constrain the subsequent action rather than being generated independently after the decision. The diagnostic step is intentionally lightweight and is not tied to a particular probabilistic classifier; it evaluates current mastery and behavioural indicators against the conditions associated with the three categories. If multiple categories are plausible, the framework selects the category ranked highest by the configured diagnostic rules.

The diagnostic space is deliberately limited to three interpretable categories. Prior cognitive-diagnostic work has shown that student difficulties can be represented using a small number of understandable categories rather than a single undifferentiated score [15]. FEAAF adopts that general design principle, while the specific categories used here are part of the proposed framework rather than being claimed as directly adopted from that study. A larger category space could support finer-grained diagnosis, but this is left for future work.

3.4 Selecting an Action Consistent with the Diagnosis

After a diagnosis is identified, FEAAF restricts the available actions to those appropriate for that diagnosis rather than allowing the policy to choose freely from every possible intervention. For a time-pressure-related diagnosis, for example, the system may reduce time pressure or continue without intervention, while reducing content difficulty would not belong to the admissible set for that diagnosis. Within the admissible action set, the framework can select an action according to application-specific utility and intervention cost. If diagnostic confidence is insufficient, the framework can choose not to intervene rather than commit to an unsupported diagnosis. At the conceptual level, a knowledge-gap diagnosis may permit remediation or additional practice; a time-pressure diagnosis may permit reduced time pressure or continuation; and a disengagement diagnosis may permit an engagement-oriented intervention or continuation. The exact mapping is application-dependent and is not empirically validated in this paper. This restriction produces the framework's central guarantee. Because an action is selected only from the set associated with the diagnosed explanation, the selected action is constrained by construction to remain consistent with the explanation shown to the learner. This is a property of policy construction, not evidence that the diagnosis is correct, that the selected action is pedagogically optimal, or that learning outcomes improve; those questions require empirical evaluation.

3.5 Worked Example of FEAAF Operation

This subsection presents a hypothetical worked example, together with two brief contrasting cases, to show step by step how CKBM, the diagnostic mechanism in Sec. 3.3, and ECAP in Sec. 3.4 operate for one learner interaction. The example is a conceptual demonstration rather than an experiment or empirical result: no learner data were collected and no implementation was used to generate the values.

Consider a hypothetical learner working on Java loops exercises who answers a question incorrectly. The learner state $S_t = (M_t, B_t)$ reflects adequate mastery for loops, based on prior correct answers on similar items; a response time substantially longer than the learner's typical pace; and stable engagement. Under Sec. 3.3, low mastery is not supported, since mastery is adequate, and declining engagement is not supported, since engagement is stable; adequate mastery with an unusually long response time matches the condition for time-pressure-related difficulty, so the framework assigns this diagnosis.

FEAAF presents the explanation “Your performance appears to be affected by time pressure.” ECAP (Sec. 3.4) therefore restricts the admissible actions to reducing time pressure for the next question or continuing without intervention. The framework selects the time-pressure reduction action. It does not reduce the material's difficulty or assign remediation because those actions are not in the admissible set for this diagnosis. Thus the explanation presented to the learner and the action selected by the policy remain consistent by construction.

Table 2 summarizes the example as a processing pipeline. The states and outputs are qualitative and illustrative; no numerical thresholds, probabilities, accuracy values, or measured learner data are introduced.

Step	Input / State	FEAAF Processing	Output
1. Interaction	Incorrect answer to a loops question	Logged into the shared stream (Sec. 3.2)	Response: incorrect, with timing
2. Mastery estimate	Prior correct answers on similar items	Per-skill mastery update (Sec. 3.2)	Mastery: adequate
3. Behavioural indicators	Response time; recent engagement	Behavioural update from timing and correctness (Sec. 3.2)	Timing: unusually long; engagement: stable
4. Diagnosis	$S_t = (M_t, B_t)$	Evidence vs. three diagnostic conditions (Sec. 3.3)	Time-pressure-related difficulty
5. Explanation	Selected diagnosis	Diagnosis mapped to learner-facing statement	“Affected by time pressure”
6. Admissible actions	Diagnosis	Action set restricted by ECAP (Sec. 3.4)	{reduce time pressure, continue}
7. Selected action	Admissible set	Action chosen from set	Reduce time pressure
8. Consistency	Explanation + action	Checked against admissible set	Consistent by construction (Sec. 3.7)

Table 2. Hypothetical worked example of one FEAAF cycle. Values are illustrative qualitative states, not measured or empirical results.

The same observable event can lead to different diagnostic explanations depending on the broader learner state. Case A represents a knowledge gap: low mastery, ordinary timing, and stable engagement, leading

to an explanation that additional practice may be needed and an admissible remediation action. Case B represents disengagement: mastery is not clearly low, several errors occur, and engagement declines, leading to an explanation referring to reduced engagement and an admissible engagement-oriented intervention or continuation.

These cases illustrate that the same surface-level event does not necessarily imply the same instructional response. FEAAF first evaluates the combined learner state and identifies a diagnosis, then restricts the available action set accordingly. This is why CKBM and ECAP are used together. The construction guarantees consistency between the diagnosis used for explanation and the admissible action; it does not establish that the diagnosis is objectively correct or that the selected intervention improves learning. Since the framework targets on-device deployment, this sequence can conceptually execute as a lightweight learner-state update, diagnosis, action restriction, and selection process on the learner's device. No Android implementation, latency measurement, or performance result is claimed here.

3.6 Keeping the Learner-State Update Lightweight

Because FEAAF is intended for on-device use, the learner-state update in Sec. 3.2 is deliberately lightweight. Maintaining every possible joint combination of skill states becomes increasingly expensive as the number of skills grows. FEAAF instead updates each relevant skill's mastery estimate from the learner's response and maintains behavioural indicators separately using interaction evidence such as response timing and correctness. This keeps each update tied to the skills affected by the current interaction rather than to every skill in the system. The approach is a standard simplification for lightweight learner-state updating, not a new computational technique claimed by this paper. The specific FEAAF design choice is to retain mastery and behaviour within one shared learner state while applying this lightweight update strategy because of the intended on-device setting. Figure 2 summarizes the process. The difference between this approximation and a full joint model remains an empirical question for future work.

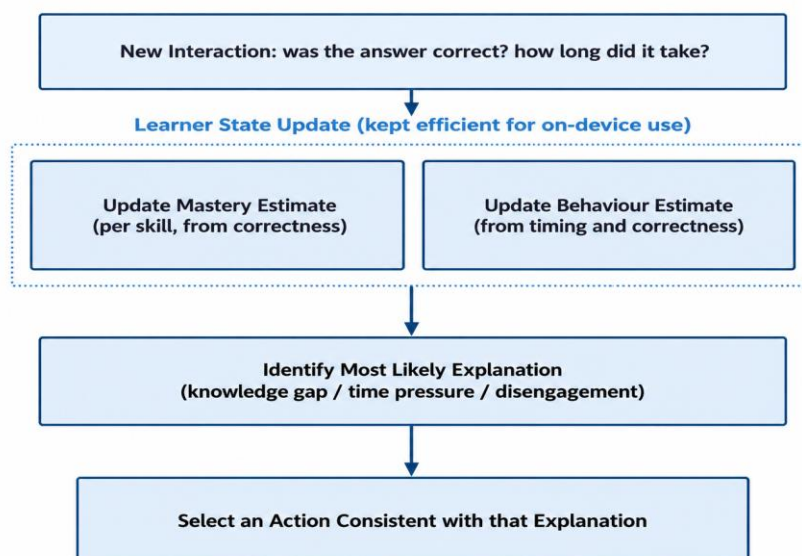


Figure 2. The learner-state update: mastery and behavioural estimates are updated separately from a shared interaction, keeping the process lightweight for on-device use before the system identifies an explanation and selects a consistent action.

3.7 Explanation-Action Faithfulness: A Structural Property

The property described in Sec. 3.4 is called Explanation-Action Faithfulness: when a system presents an explanation naming a particular diagnosis, the action it executes belongs to the set of actions associated with that diagnosis. The property does not depend on a particular learner representation. FEAAF is one framework for realizing it through CKBM and its diagnostic mechanism, but the same property could in principle be examined in other adaptive systems.

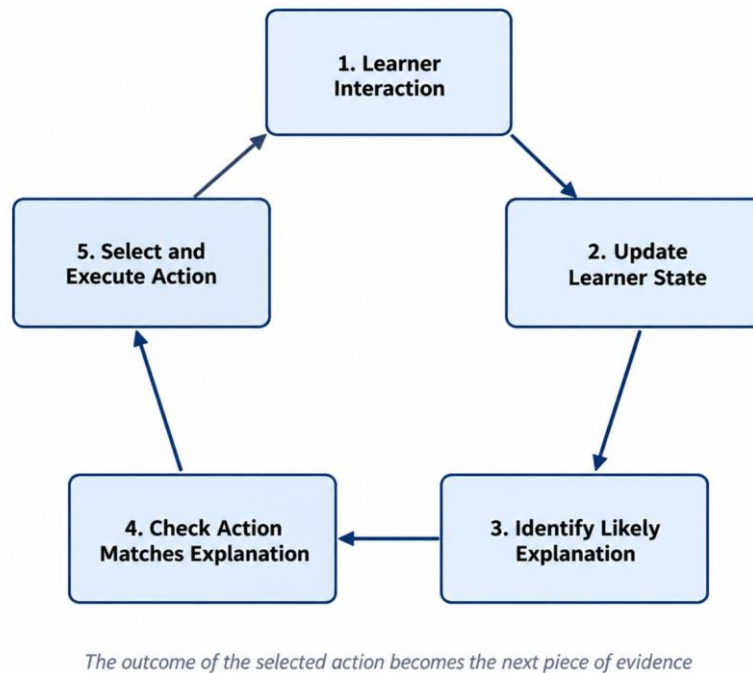


Figure 3. FEAAF's continuous cycle: each learner interaction updates the learner state, informs the likely explanation, and constrains the action selected; the outcome of that action then becomes part of the evidence for the next interaction.

Because ECAP (Sec. 3.4) selects actions only from the set associated with the diagnosed explanation, the action executed by FEAAF is consistent with the explanation it presents by construction. This follows from restricting the admissible action set before selecting an action; it is therefore a structural property of the policy rather than a claim about diagnostic correctness. The property does not establish that the diagnosis is correct, that the admissible actions are pedagogically optimal, that an intervention improves learning, that learners or educators will trust the explanation, or that FEAAF outperforms systems such as BKT or DKT. Its narrower claim is that the explanation shown and the action selected cannot contradict each other under the policy design.

4. Comparative Analysis

Table 1 compares FEAAF with the systems discussed in Sec. 2 using the specific property introduced by this paper: whether consistency between an explanation and the subsequent action is guaranteed by design. Systems such as BKT, DKT, and IRT-based testing do not provide the same explanation-action property because they do not use an equivalent explanation mechanism. For systems that do provide explanations,

such as rule-based, graph-based, or attribution-based approaches, the reviewed literature does not identify an explicit constraint that links the explanation to the action-selection step. This comparison is limited to the literature reviewed in this paper and is not claimed to be exhaustive.

System	Mastery Signal	Behavioural Signal	Explanation	Action Matches Explanation?
BKT / DKT [2],[3]	Yes	No	None / opaque (DKT)	N/A
IRT-Based Testing	Yes	No	None	N/A
Attribution-Based EDM [4]	No	No	Post-hoc attribution	Not guaranteed
Cognitive Tutors [10]	Yes	No	Rule trace	Not guaranteed
Knowledge-Graph Tutoring [11]	Yes	No	Opaque embedding	Not guaranteed
Rule-Based ITS [12]	Partial	No	Rule justification	Not guaranteed
FEAAF (proposed)	Yes	Yes	Diagnosis-based	Guaranteed by design

Table 1. Comparison of FEAAF against the systems discussed in Sec. 2, focused on whether the action taken is guaranteed to remain consistent with the explanation shown.

5. Discussion

For educational data mining and learning analytics, FEAAF provides a concrete criterion—explanation-action consistency—that future systems can evaluate independently of their learner representation. For intelligent tutoring systems, CKBM provides a way to track behavioural information alongside mastery while retaining an interpretable state representation. For mobile and on-device AI, the lightweight per-skill update described in Sec. 3.6 illustrates how the combined learner state can be designed around Android hardware constraints rather than treating local deployment as an afterthought.

6. Limitations and Future Work

This paper presents a conceptual and formal framework rather than an implemented and evaluated system, so several questions remain open for empirical research. Updating mastery and behaviour separately is intended to reduce computational demands, but the resulting trade-off with a full joint model has not been measured. The three-category diagnostic space is deliberately limited for interpretability, and expanding it may introduce additional ambiguity or computational cost; this also requires empirical study. The mapping from diagnoses to admissible actions is proposed conceptually, and validating whether those mappings reflect genuine pedagogical relationships rather than plausible heuristics would require educator input and empirical testing. Future work is therefore described as a staged research programme. Stage 1 is the present conceptual FEAAF framework. Stage 2 would implement CKBM and ECAP on Android.

Stage 3 would test whether explanation-action consistency holds in a running implementation under controlled conditions. Stage 4 would compare CKBM mastery estimates with BKT and DKT using public knowledge-tracing datasets. Stage 5 would examine how learners and educators interpret and trust the faithfulness property. Stage 6 would evaluate whether the resulting interventions improve learning outcomes relative to appropriate comparison systems. None of these stages has been carried out in this paper; they are proposed as future research rather than reported results.

7. Conclusion

This paper introduced Explanation-Action Faithfulness, the requirement that an adaptive educational system keep its selected action consistent with the explanation presented to the learner, and proposed FEAAF as a framework for achieving this through CKBM and ECAP. CKBM combines mastery and behavioural information in a shared learner state, while the diagnostic step identifies a likely explanation and ECAP restricts subsequent actions to those compatible with that explanation. Within the literature reviewed for this paper, no comparable approach was identified that explicitly guarantees this consistency at the action-selection stage. The lightweight per-skill state update is motivated by the constraints of Android on-device deployment, showing how the deployment setting can influence the framework design. The paper does not claim that explanation-action consistency establishes diagnostic correctness or pedagogical effectiveness; it isolates consistency between what the system explains and what it does as a distinct structural property that can be tested empirically in future work.

REFERENCES:

- [1] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?," in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), 2020, pp. 4198–4205, doi: 10.18653/v1/2020.acl-main.386.
- [2] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Modeling and User-Adapted Interaction*, vol. 4, no. 4, pp. 253–278, 1994, doi: 10.1007/BF01099821.
- [3] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," in *Advances in Neural Information Processing Systems 28 (NeurIPS 2015)*, 2015, pp. 505–513.
- [4] Rahul and R. Katarya, "Shapley explainable deep learning based knowledge distillation framework for student's performance prediction," in Proc. 2024 2nd Int. Conf. on Data Science and Information System (ICDSIS), 2024, pp. 1–8, doi: 10.1109/ICDSIS61070.2024.10594386.
- [5] A. Khan and S. K. Ghosh, "Student performance analysis and prediction in classroom learning: A review of educational data mining studies," *Education and Information Technologies*, vol. 26, no. 1, pp. 205–240, 2021, doi: 10.1007/s10639-020-10230-3.
- [6] A. Khan and S. K. Ghosh, "Data mining based analysis to explore the effect of teaching on student performance," *Education and Information Technologies*, vol. 23, no. 4, pp. 1677–1697, 2018, doi: 10.1007/s10639-017-9685-z.
- [7] S. Dhara, S. Chatterjee, R. Chaudhuri, A. Goswami, and S. K. Ghosh, "Artificial intelligence in assessment of students' performance," in *Artificial Intelligence in Higher Education*, Boca Raton, FL: CRC Press, 2022, pp. 153–167, doi: 10.1201/9781003184157-8.

- [8] Rahul and R. Katarya, "A systematic review on predicting the performance of students in higher education in offline mode using machine learning techniques," *Wireless Personal Communications*, vol. 133, pp. 1643–1674, 2023.
- [9] N. Nithiyanandam et al., "Artificial intelligence assisted student learning and performance analysis using instructor evaluation model," in *Proc. 2022 3rd Int. Conf. on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2022, pp. 1555–1561, doi: 10.1109/ICESC54411.2022.9885462.
- [10] J. R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier, "Cognitive tutors: Lessons learned," *The Journal of the Learning Sciences*, vol. 4, no. 2, pp. 167–207, 1995, doi: 10.1207/s15327809jls0402_2.
- [11] H. Nakagawa, Y. Iwasawa, and Y. Matsuo, "Graph-based knowledge tracing: Modeling student proficiency using graph neural network," in *Proc. 2019 IEEE/WIC/ACM Int. Conf. on Web Intelligence (WI)*, 2019, pp. 156–163, doi: 10.1145/3350546.3352513.
- [12] K. VanLehn, "The behavior of tutoring systems," *International Journal of Artificial Intelligence in Education*, vol. 16, no. 3, pp. 227–265, 2006.
- [13] R. Pekrun, "The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice," *Educational Psychology Review*, vol. 18, no. 4, pp. 315–341, 2006, doi: 10.1007/s10648-006-9029-9.
- [14] M. Tompala, N. Padhy, R. Panigrahi, and S. P. Patro, "Evaluating regression models for student grade prediction with a focus on interpretability and accuracy," in *Intelligent Computing and Communication (ICICC 2025)*, *Lecture Notes in Networks and Systems*, vol. 1839, Cham, Switzerland: Springer, 2026, doi: 10.1007/978-3-032-18349-1_30.
- [15] M. B. Garcia, "Profiling the skill mastery of introductory programming students: A cognitive diagnostic modeling approach," *Education and Information Technologies*, 2024, doi: 10.1007/s10639-024-13039-6.