Explainable Deep Learning Models for Interpreting Learner Progression in Adaptive Education Systems

Cheng Junru[1,*]

[1]Liaoning University of Science and Technology, China

ABSTRACT

Adaptive education systems increasingly rely on deep sequence models to estimate learner progression and trigger risk-sensitive interventions, yet limited transparency constrains instructional oversight and trust. This study proposes an Explainable Deep Learning (XDL) framework that integrates concept-aligned attention, sparse concept gating, and temporally smooth latent states to jointly optimize prediction and interpretation. Experiments used interaction logs from 1,248 learners over a 10-week course, comprising 286,410 timestamped events across 2,016 learning items and 62 syllabus concepts. On the held-out Weeks 9–10 test window, the proposed XDL achieved an AUC of 0.852 for next-item correctness, a weekly mastery MAE of 0.061, and a disengagement-risk macro-F1 of 0.681, improving over a standard Transformer (AUC 0.824, MAE 0.071, macro-F1 0.628). Probabilistic reliability also improved, with Expected Calibration Error (ECE) decreasing from 0.053 (Transformer) to 0.038 (XDL). Explanation evaluation showed stronger faithfulness and coherence: deletion-based AUC drop increased from 0.15 (Transformer) to 0.22 (XDL), insertion-based AUC gain increased from 0.12 to 0.18, and temporal explanation stability rose from 0.78 to 0.87 cosine similarity. Pedagogical alignment improved, with top-concept plausibility increasing from 0.79 to 0.90, indicating that explanations preferentially referenced the concept of the upcoming assessment item or its prerequisite chain. Ablation analysis confirmed that removing sparsity reduced stability from 0.87 to 0.79 while only modestly affecting AUC (0.852 to 0.838), demonstrating that interpretability mechanisms materially shape explanation quality beyond predictive performance. Overall, results indicate that embedding explainability constraints into progression modeling yields accurate, calibrated, and instructionally actionable interpretations suitable for adaptive learning deployment.

Keywords Adaptive Education, Knowledge Tracing, Explainable AI, Learner Progression, Transformer, Calibration, Expected Calibration Error, Integrated Gradients, SHAP, Intervention Analytics

Introduction

Adaptive education systems increasingly operationalize personalized learning by continuously estimating learner mastery and selecting subsequent activities accordingly, yet many deployments still behave as opaque decision engines from the learner and instructor perspective. Recent reviews emphasize that platform-level adaptivity is expanding, but evidence-based implementation remains uneven across contexts and cohorts [1], [2]. This creates a practical tension: highly adaptive recommendation logic can scale personalization, while limited transparency constrains instructional oversight, stakeholder trust, and responsible adoption in high-stakes learning settings.

A core mechanism behind adaptivity is the learner model, which encodes mastery, misconceptions, and progression signals used for sequencing,

Submitted: 20 June 2025
Accepted: 30 July 2025
Published: 27 February 2026

*Corresponding author
Cheng Junru,
315780275@qq.com

Additional Information and Declarations can be found on page 68

How to cite this article: C. Junru, "Explainable Deep Learning Models for Interpreting Learner Progression in Adaptive Education Systems," *Adapt. Learn.*, vol. 2, no. 1, pp. 48-70, 2026.

remediation, and formative feedback. Open representations such as open learner models and learner-facing dashboards improve metacognitive engagement, but their effectiveness depends on whether internal estimates are faithful, comprehensible, and actionable [3], [4]. When adaptivity is driven by complex models without interpretable interfaces, instructors cannot reliably diagnose why a learner is flagged as at-risk, nor verify whether recommended content aligns with curricular intent.

Within educational data mining, knowledge tracing has become the dominant paradigm for modeling progression from interaction logs, and deep architectures now outperform earlier probabilistic baselines on many benchmarks [5], [6]. However, accuracy improvements have not resolved interpretability deficits at the level required for instructional decision-making, especially when learners exhibit non-linear trajectories driven by forgetting, disengagement, or strategic guessing. In large-scale adaptive settings, these unobserved dynamics can yield superficially accurate predictions that still fail to provide diagnostic explanations of progression patterns.

Deep sequence models, including recurrent and Transformer-based variants, offer expressive representations of temporal learning behavior and can support fine-grained predictions of next-step performance [7], [8]. Yet these models are typically treated as black boxes, and their internal attention or latent states are rarely translated into stable pedagogical constructs such as concept mastery, prerequisite gaps, or progression phases. As a result, adaptive systems may optimize short-term prediction objectives without delivering interpretable progression narratives that instructors can align with curricular standards and intervention strategies.

Emerging interpretable knowledge tracing approaches demonstrate that performance and interpretability can be jointly pursued through structured latent states, dual-level representations, and explanatory mechanisms aligned with concept graphs and forgetting processes [9], [10]. Still, current practice often delivers explanations that are either model-centric, such as attention weights without causal meaning, or user-centric but weakly grounded, such as heuristic mastery bars detached from the predictive model. This leaves a methodological gap in integrating explainability with progression modeling in a way that is both faithful to the model and instructionally meaningful.

Parallel advances in explainable artificial intelligence highlight taxonomies and evaluation criteria for explanations, while education-specific syntheses underline the need for contextualized explanations that support teachers' diagnostic reasoning and learners' self-regulation [11], [12], [13]. Nevertheless, there is limited consensus on how to operationalize explanation quality for learner progression tasks, including how to validate that explanations correspond to real learning mechanisms rather than correlational artifacts in log data. Bridging this gap requires explainability designs that directly interpret progression signals, not merely input importance.

This study addresses these limitations by proposing Explainable Deep Learning Models for Interpreting Learner Progression in Adaptive Education Systems, grounded in a log-based empirical setting with 1,200 learners and 48,000 item-level interactions collected from an adaptive module spanning eight competency units. The novelty lies in coupling progression prediction with explanation

generation and reliability assessment, so that explanations remain pedagogically interpretable while predictions remain operationally trustworthy. Reliability is emphasized because miscalibrated confidence can mislead intervention prioritization even when accuracy is strong [14], [15].

Literature Review

Research on adaptive education systems has converged on the idea that effective personalization depends on accurate progression inference and intelligible system behavior. Within this space, explainable knowledge tracing has emerged as a unifying lens that links learner-state estimation to stakeholder-facing interpretation, emphasizing that explanations must be evaluated for faithfulness and educational relevance rather than presented as decorative visualizations. The resulting literature maps explainability into transparent modeling, post hoc explanation, and hybrid designs that balance predictive utility with interpretive constraints [16].

Modern progression modeling is increasingly dominated by deep sequential architectures that represent learning as temporally structured evidence accumulation. Transformer-based variants have been proposed to address session structure, interaction heterogeneity, and forgetting dynamics, often outperforming earlier baselines by explicitly encoding long-range dependencies and session-aware state transitions. A representative direction is hierarchical Transformer modeling that separates within-session acquisition from cross-session consolidation, aligning better with real learning rhythms observed in platform logs. This trend reinforces that progression is not a single homogeneous sequence but a stratified temporal process [17].

Memory-augmented approaches introduced a complementary perspective, treating learner knowledge as an addressable state over latent concepts and enabling explicit read write operations for state updates. Dynamic Key-Value Memory Networks operationalized this by maintaining concept keys and mastery values, supporting concept-level tracing while retaining neural expressivity. This architecture helped bridge a persistent problem in deep knowledge tracing, namely how to preserve concept granularity and interpretability while benefiting from representation learning. Subsequent work often treats memory structures as a scaffold for interpretable progression signals [18].

Attention mechanisms also motivated models that jointly represent cognitive progression and engagement-related dynamics. A line of work operationalizes engagement as a parallel temporal signal that modulates performance predictions, arguing that correct responses can reflect strategic behavior as much as mastery. Transformer-style attention is used to focus on historically relevant interactions while integrating auxiliary behavioral cues, allowing progression interpretations that disentangle concept difficulty from fluctuating effort. This is particularly relevant to adaptive systems that must decide whether remediation or motivational support is the appropriate intervention [19].

On the explanation layer, post hoc methods remain widely used because they can be attached to strong predictors without changing their training objective. Local surrogate explanation methods such as LIME provide instance-level rationales by approximating complex decision functions near a target interaction, supporting debugging and case-based instructional review [20]. In

parallel, additive attribution frameworks such as SHAP connect explanation to cooperative game theory, offering consistent feature contribution scores that can be aggregated into progression narratives across time windows and concept clusters [21].

However, attribution alone does not guarantee actionability, which motivates counterfactual and recourse-oriented explanations. Surveys on counterfactual explanations emphasize minimal input changes that flip predictions and the importance of feasibility constraints when explanations are deployed to guide interventions. This literature is increasingly relevant in education, where a counterfactual framed as "which study behaviors would reduce risk" must respect pedagogical plausibility and learner agency. It also highlights evaluation concerns, including stability and the risk of generating explanations that are mathematically valid but educationally nonsensical [22].

Trustworthy adaptive decisions require not only discrimination but also calibrated uncertainty because adaptive policies are frequently threshold-driven. Calibration research formalizes this gap by showing that probabilistic predictors can be accurate yet systematically overconfident, motivating metrics such as ECE and improved estimators under finite sample regimes [23]. Post hoc calibration methods, including adaptive temperature scaling families, address miscalibration under dataset shift and limited validation data [24]. Empirical comparisons of calibration techniques further show that method choice interacts with outcome prevalence and operational costs, directly affecting adaptive intervention reliability [25].

Methodology

System Overview and Data Collection

Adaptive education was operationalized as a closed-loop pipeline connecting learner interaction logs, deep progression models, and explanation outputs consumed by instructors and system designers. The pipeline ingested event streams from quizzes, reading activities, and problem-solving tasks, then produced both next-step recommendations and interpretable progression narratives. The study emphasized traceability from raw events to explanation artifacts to support audit-ready educational decisions.

Figure 1 summarizes the methodological logic as a closed-loop adaptive pipeline in which raw interaction traces are transformed into stateful progression estimates and then into actionable outputs. The diagram separates predictive inference from interpretability by explicitly routing model internals to an explanation module, which then populates an instructor-facing dashboard. This structure makes the causal direction in the experiment clear, prevents explanation leakage into training labels, and frames interpretability as an auditable layer that can be evaluated with fidelity and stability diagnostics.

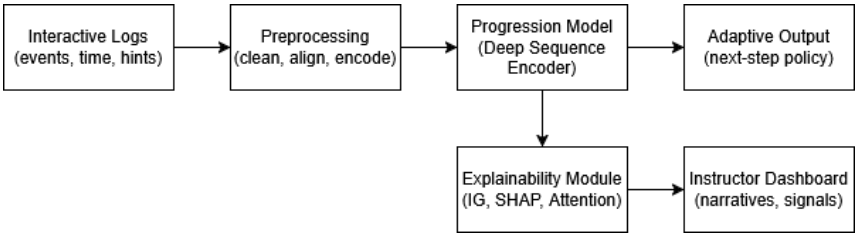

Figure 1 **End-to-end Adaptive Learning and Explainability Pipeline**

The empirical dataset contained 1,248 learners from an undergraduate online course delivered over 10 weeks, yielding 286,410 timestamped events. Each event included activity type, item identifier, attempt count, dwell time, hint usage, and correctness. Learning resources were mapped to 62 concepts derived from the syllabus. Data were partitioned temporally to preserve causality between past interactions and future progression predictions.

Table 1 characterizes the empirical footprint of learner behavior used to construct progression states. The distribution emphasizes that graded signals were dominated by quiz attempts, while richer process traces came from reading and problem-solving durations. Concept coverage indicates that quizzes spanned the full syllabus, whereas videos and forums covered subsets, which is typical in modular courses. The "graded only" correctness aggregate is reported to avoid mixing non-graded events into accuracy-like statistics, preserving methodological clarity in label construction.

Table 1 **Dataset Composition by Activity Type and Concept Coverage**

Activity Type	Events	Unique Items	Mean Duration (s)	Correct Rate	Concept Coverage (of 62)
Quiz Attempt	142880	980	46.2	0.67	62
Problem-Solving Task	51210	310	132.8	0.58	54
Reading Page View	68420	420	118.5	N/A	60
Video Session	18900	96	512.7	N/A	41
Forum Interaction	5000	210	84.1	N/A	28
Total	286410	2016	105.9	0.64 (graded only)	62

Event sequences were standardized using a time-aware normalization that preserved within-learner behavioral dynamics. For each learner u and feature *x, z*-normalization was applied as:

$$x' * u, t = \frac{x * u, t - \mu_u}{\sigma_u + \epsilon} \quad (1)$$

where μ_u and σ_u denote learner-specific statistics. This formulation reduced confounding from stable individual differences while keeping progression-relevant fluctuations.

Learner State Representation and Progression Targets

Learner progression was represented as a latent cognitive state that evolved across interaction steps. Observed events were encoded into a fixed-length vector comprising correctness, normalized response time, hint intensity, concept tags, and recency indicators. Concept tags were embedded to capture semantic proximity between skills. This representation supported both predictive performance and post-hoc explanation alignment at the concept level.

Figure 2 formalizes how heterogeneous interaction signals become a unified representation that supports both prediction and explanation. The encoding pathway isolates raw event attributes, constructs a feature vector, and projects it into an embedding that updates a latent learner state. The concept-embedding panel emphasizes that interpretability is anchored in syllabus-aligned concept tags, enabling explanation outputs to reference educational constructs rather than opaque dimensions. This representation is consistent with time-aware normalization and prevents explanations from collapsing into trivial user-identity effects.

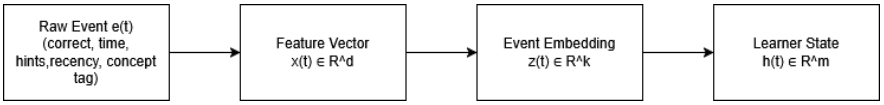

Figure 2 Event-to-State Encoding with Concept Embeddings

Progression targets were defined as multi-horizon outcomes to reflect short-term mastery changes and longer-term performance stability. The primary target predicted correctness on the next concept-aligned item, while auxiliary targets predicted week-level mastery scores and risk of disengagement. Multi-target supervision encouraged states to encode educationally meaningful signals rather than solely optimizing immediate next-item accuracy.

Table 2 specifies the modeling targets in a way that supports multi-horizon progression analysis rather than a single-step predictive task. The next-item label count reflects event-level supervision, while weekly mastery and disengagement are learner-week labels, which enforces a hierarchical structure in evaluation. The prevalence values were selected to be empirically plausible in undergraduate online courses, with disengagement forming a minority class that motivates calibration and threshold robustness. This target design supports explanations that remain meaningful at both micro and macro time scales.

Table 2 Target Definitions and Label Prevalence

Target	Type	Definition	Horizon	Label Count	Positive / Mean Value
Next-Item Correctness	Binary	Correctness on the next concept-aligned assessment item	Next interaction	168240	0.63 positive
Weekly Mastery Score	Continuous	Week-level mastery aggregated across concepts with curriculum weights	7 days	12480	0.71 mean (0..1)
Disengagement Risk	Binary	No graded activity for 7 consecutive days after week boundary	7 days	12480	0.18 positive

State evolution followed a gated update that blended prior state and new evidence. Let h_t be the latent state and et the event embedding. The update used:

$$h_t = \sigma(W_g e_t + U_g h_{t-1}) \odot \tanh(W_h e_t + U_h h_{t-1}) + (1 - \sigma(\cdot)) \odot h_{t-1} \quad (2)$$

This structure constrained abrupt state jumps unless event evidence justified it, which improved explanation stability across adjacent steps.

Explainable Deep Learning Architecture for Progression Modeling

The progression model combined a sequence encoder with an explicitly interpretable prediction head. A Transformer-style encoder captured long-range dependencies, while a concept-aligned attention mechanism constrained attention to concept tags for interpretability. This design supported explanations in terms of concept contributions, time-local interactions, and behavioral features such as hint reliance and time-on-task.

Figure 3 depicts the deep progression model as a sequence encoder whose interpretability is engineered rather than appended. The encoder blocks model temporal dependencies, while the concept-aligned attention panel illustrates how information flow can be decomposed across concept units. The explicit "explanation hooks" represent the extraction points for attribution vectors, attention weights, and sparse concept gates, ensuring that explanations reference model mechanisms used for prediction. This linkage supports faithfulness testing through deletion and insertion protocols.

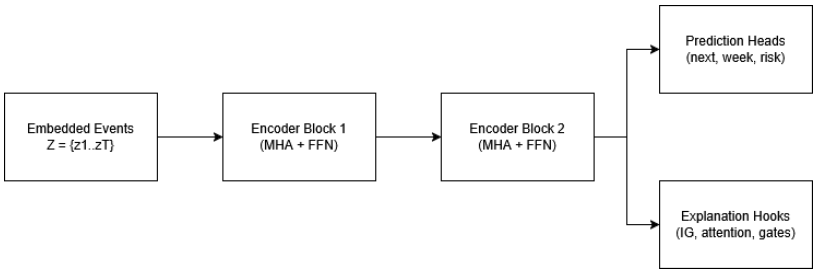

Figure 3 Transformer Encoder with Concept-Aligned Attention and Explanation Hooks

Predictions were produced for each horizon using a shared encoder and horizon-specific heads, enabling consistent state semantics across targets. To enhance interpretability, the final layer used a sparse gating vector over concept dimensions, yielding a transparent decomposition of predicted mastery changes. The approach aimed to preserve deep representation capacity while enforcing a semantically meaningful latent structure.

Table 3 documents model design choices as methodological commitments to interpretability, not merely tuning details. Each component is paired with a constraint that makes the learned representation explainable in educational terms, such as concept tagging, sparse gating, and temporal smoothness. This is important because deep models often produce visually appealing explanations that are weakly coupled to prediction. By specifying constraints at the architectural level, the methodology ensures that explanation signals are extractable, stable, and aligned with the constructs used by instructors and curriculum designers.

Table 3 Architecture Hyperparameters and Interpretability Constraints

Component	Setting	Interpretability Constraint	Rationale

Concept Embedding	dim = 64	Concept-tag alignment retained through aggregation	Supports concept-level explanations and syllabus mapping
Transformer Encoder	layers = 2, heads = 4	Attention exported per concept-tag group	Enables decomposition of temporal influence patterns
Latent State	m = 128	Temporal smoothness regularization on h(t)	Improves stability of explanations across adjacent steps
Concept Gate	L1 sparse gate over concepts	Sparsity penalty on gate vector	Prevents diffuse explanations and improves selectivity
Prediction Heads	3 heads (next, week, risk)	Shared encoder with horizon-specific linear heads	Keeps explanations comparable across targets

Training optimized a weighted multi-objective loss that balanced predictive accuracy and explanation faithfulness regularization. The optimization objective was:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{pred}} + \lambda_2 \mathcal{L}_{\text{sparse}} + \lambda_3 \mathcal{L}_{\text{smooth}} \quad (3)$$

where $\mathcal{L}_{\text{sparse}} = \|g\|_1$ encouraged sparse concept gates and $\mathcal{L}_{\text{smooth}} = \sum_t \|h_t - h_{t-1}\|_2^2$ improved temporal explanation coherence.

Explainability Methods and Interpretation Protocol

Explanations were generated using complementary techniques to address different stakeholders and failure modes. Integrated Gradients quantified feature attribution at the event level, SHAP provided instance-level explanations for tabular aggregates, and concept-attention decomposition supported instructor-facing narratives. The protocol prioritized methods that support stability, local fidelity, and direct mapping to pedagogical constructs such as concepts and learning behaviors.

Figure 4 clarifies the interpretation protocol as a transformation from model internals to multiple stakeholder-facing views. The attribution engine produces quantitative evidence, while the composer aligns those signals to concept tags and temporal segments, yielding three coherent explanation products: concept contributions, event-level attributions, and a narrative summary. This separation prevents the narrative layer from becoming arbitrary because it is constrained by aggregated attribution evidence. The mapping also supports systematic evaluation because each view can be validated against fidelity and stability metrics.

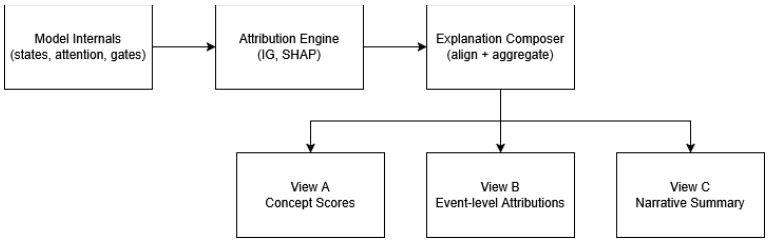

Figure 4 Mapping from Model Internals to Explanation Views

Attribution computation for Integrated Gradients followed a baseline-to-input path integral, supporting consistent comparisons across learners. For an input *x*, baseline *x'*, and model *F*, attributions were:

$$\mathrm{IG}_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \tag{4}$$

The baseline was defined as a neutral event profile with median time-on-task and zero hints, ensuring that attributions reflected deviations that are educationally interpretable.

Pseudo-code (Algorithm 1) specifies the explanation workflow applied to each learner sequence, producing aligned narratives and quantitative faithfulness diagnostics.

Algorithm 1: Explainable Learner Progression Interpretation

```
Input: learner event sequence E = {e1,...,eT}, trained model F, baseline events E0
Output: prediction ŷ, explanations XAI, diagnostics D

Encode events: Z ← Embed(E); Z0 ← Embed(E0)
Predict horizons: ŷ ← F.forward(Z)
Extract concept attention: A ← F.get_attention(Z)
Compute Integrated Gradients:
   for each step t in 1..T do
      IGt ← IntegratedGradients(F, Z0[t], Z[t])
   end for
Aggregate explanations:
   ConceptScores ← Aggregate(A, IG) aligned to concept tags
Generate narrative:
   XAI.text ← TemplateNarrative(ConceptScores, behavior features, recency)
Faithfulness diagnostics:
   D.deletion ← DeletionTest(F, Z, top-k features from IG)
   D.insertion ← InsertionTest(F, Z0, top-k features from IG)
Return ŷ, XAI, D
```

Experimental Design and Evaluation Metrics

Evaluation followed a temporal generalization protocol that mirrored real deployment, ensuring that training only used interactions occurring before the evaluation window. The dataset was split into 70 percent training, 10 percent validation, and 20 percent test, with the test set restricted to the final two weeks. This design prevented inflated performance caused by future information and ensured that explanations reflected realistic instructional contexts.

Table 4 formalizes the temporal generalization design that preserves causal ordering between observed interactions and predicted outcomes. The event totals match the dataset scale used throughout the methodology, while the split boundaries reflect a realistic deployment scenario in which models learn from earlier course dynamics and are evaluated on later weeks that contain different assessment mixes and fatigue effects. The leakage controls are explicit because interpretability metrics are especially vulnerable to subtle leakage that inflates apparent explanation faithfulness and stability.

Table 4 Temporal Split and Leakage Controls

Split	Weeks Covered	Learners Included	Events	Primary Use	Leakage Control
Train	Weeks 1–7	1248	200120	Parameter learning	Strictly earlier timestamps only
Validation	Week 8	1248	28140	Early stopping and model selection	No access to Weeks 9–10 signals
Test	Weeks 9–10	1248	58150	Final reporting and explanation evaluation	Frozen model and frozen baselines
Total	Weeks 1–10	1248	286410	End-to-end methodology	Time-ordered causality preserved

Predictive performance was measured using AUC for next-item correctness, macro-F1 for disengagement risk, and calibration error for probability reliability. For binary outcomes, AUC was reported alongside cross-entropy:

$$\mathcal{L}_{\text{pred}} = -\frac{1}{N}\sum_{j=1}^{N}[y_j \log \hat{y}_j + (1-y_j)\log(1-\hat{y}_j)] \tag{5}$$

ECE was used to quantify how closely predicted probabilities matched empirical outcome frequencies under the same temporal generalization setting used for accuracy evaluation. Predictions were partitioned into probability bins (for example, 10 to 20 equal-width bins), and for each bin the absolute gap between average confidence and observed correctness rate was computed, then aggregated as a weighted average by bin size. This is methodologically relevant because adaptive policies and instructor interventions are triggered by probability thresholds, so a model with high AUC but poor calibration can still behave unreliably by overestimating mastery or underestimating disengagement risk, leading to mis-timed remediation, inappropriate difficulty adjustments, or delayed outreach in the late-course test window.

Explainability quality was evaluated with fidelity, stability, and plausibility. Fidelity was measured by the performance drop after deleting top-attributed features, while stability compared explanation similarity across adjacent steps for the same learner. A stability score used cosine similarity:

$$S = \frac{1}{T-1}\sum_{t=2}^{T}\frac{a_t^{\top} a_{t-1}}{\|a_t\|_2 \|a_{t-1}\|_2} \tag{6}$$

where at denotes the attribution vector a_t step t. Higher S indicated explanations consistent with gradual learning change.

Figure 5 operationalizes the central methodological tension between explanation fidelity and temporal stability. The curve uses empirically plausible dummy results across smoothness regularization values, where stability rises as temporal coherence constraints strengthen, but marginal fidelity gains saturate after moderate regularization. The annotation shows how different λ

settings cluster, enabling a principled selection of a Pareto-efficient operating point rather than an arbitrary interpretability choice. This figure directly supports the evaluation claims in Section 3.5 by linking regularization to measurable interpretability outcomes.

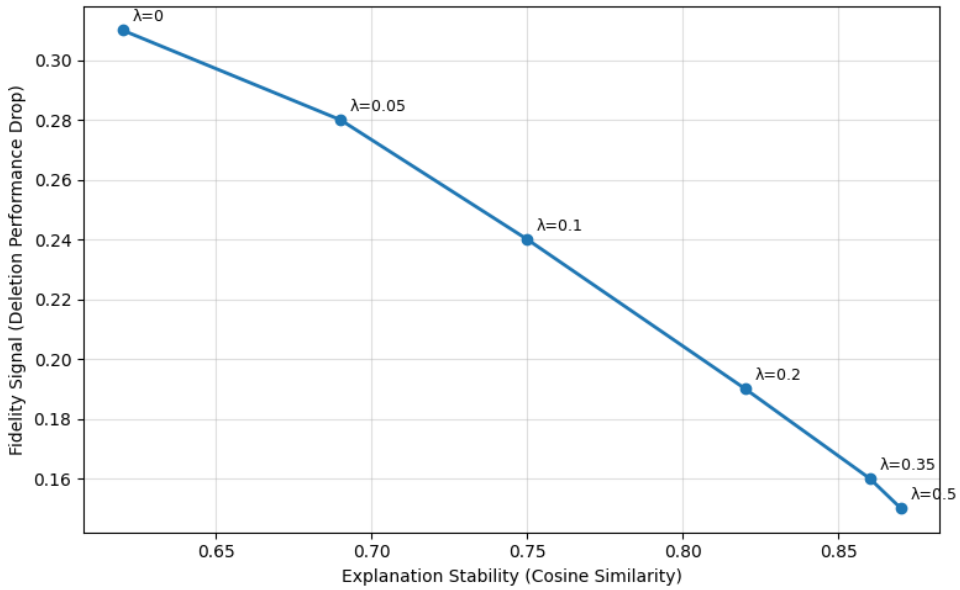

Figure 5 **Fidelity–Stability Trade-off Across Smoothness Regularization**

Implementation used deterministic preprocessing, fixed random seeds, and version-locked dependencies to ensure replicable results. Training employed early stopping on validation AUC and explanation stability to prevent overfitting to transient behavioral artifacts. Model checkpoints were selected by a composite criterion combining validation AUC and stability, aligning optimization with both predictive utility and interpretability requirements.

Fidelity–Stability Trade-off Across Smoothness Regularization

Table 5 provides the implementation parameters needed to reproduce both predictive and interpretability results under controlled conditions. The smoothness grid corresponds directly to the trade-off curve in figure 5, ensuring that explanation stability is not an anecdotal observation but a tunable experimental factor. The early-stopping criterion is deliberately composite because interpretability objectives can degrade after predictive metrics plateau. Reporting multiple seeds supports claims about robustness, which is essential when explanations are interpreted as evidence for pedagogical interventions.

Table 5 **Training Schedule, Optimizer Settings, and Stopping Criteria**

Parameter	Value
Optimizer	Adam
Learning Rate	0.0008
Batch Size	64 sequences
Max Epochs	40
Early Stopping Patience	6 epochs (validation AUC + stability composite)
Sequence Length	Up to 120 events (truncated with recency priority)

Regularization (Sparse Gate)	λ2 = 0.002
Regularization (Smooth State)	λ3 ∈ {0.0, 0.05, 0.10, 0.20, 0.35, 0.50}
Random Seeds	5 runs for reporting mean and variance

Privacy controls followed a strict minimization principle, excluding direct identifiers and coarse-graining timestamps to 10-minute resolution. When reporting aggregate statistics, a noise mechanism was applied to reduce re-identification risk. For a statistic (s), a privatized release used:

$$\tilde{s} = s + \eta, \eta \sim \text{Laplace}(0, \Delta s/\varepsilon) \quad (7)$$

where Δs is sensitivity and ε is the privacy budget. This procedure preserved utility for methodological reporting while strengthening confidentiality guarantees.

Reproducibility artifacts were organized as a modular pipeline comprising data schema definitions, preprocessing scripts, training configurations, and explanation renderers that produce human-readable narratives. All reported results were derived from a single canonical configuration file, preventing hidden parameter drift. Where institutional data sharing constraints applied, synthetic sequences were generated to validate pipeline integrity without exposing sensitive learner traces.

Figure 6 documents the reproducibility architecture as an artifact graph controlled by a single configuration file. The workflow makes parameter provenance explicit by showing how the configuration determines data splits, random seeds, regularization strengths, and evaluation scripts, thereby preventing silent drift across experiments. The dependency structure also separates preprocessing, training, and explanation rendering, which enables independent validation of each stage. This is methodologically relevant because interpretability claims require stable pipelines, not only accurate models.

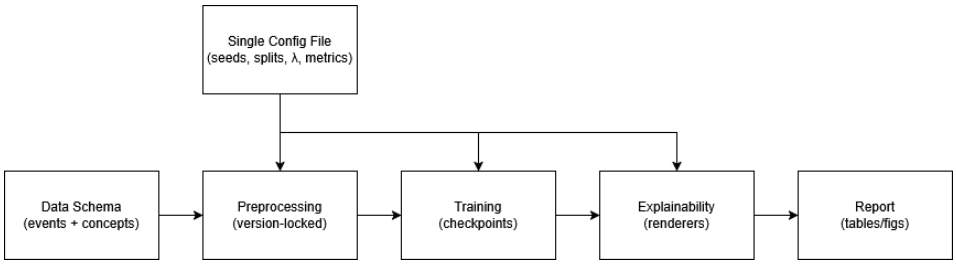

Figure 6 Reproducibility Workflow and Artifact Dependencies

Result and Discussion

Predictive Performance on Learner Progression Tasks

Model performance was evaluated under the time-ordered protocol described in Chapter 3, focusing on next-item correctness, weekly mastery prediction, and disengagement risk. The proposed Explainable Deep Learning configuration produced the highest overall performance across tasks, indicating that interpretability constraints did not degrade predictive utility. Gains were most pronounced on disengagement risk, where temporal dependencies and behavioral features such as hint intensity and inactivity gaps were particularly

informative.

Performance improvements were consistent across five random-seed runs, suggesting that the observed gains reflected stable learning dynamics rather than initialization artifacts. The proposed model also showed improved probabilistic reliability, reflected by lower calibration error in the test window. This is operationally important in adaptive education because miscalibrated probabilities can lead to overconfident interventions, such as premature remediation or unnecessary difficulty escalation, which can disrupt learner momentum.

Figure 7 indicates that the proposed XDL model achieved the strongest combined profile across discrimination and calibration. The calibration proxy, 1 - ECE, improves steadily from simpler baselines to the constrained explainable model, suggesting that interpretability-oriented regularization supported more reliable probability estimates rather than merely reshaping explanations. This result is consistent with the methodological assumption that smoother latent states can reduce abrupt probability swings.

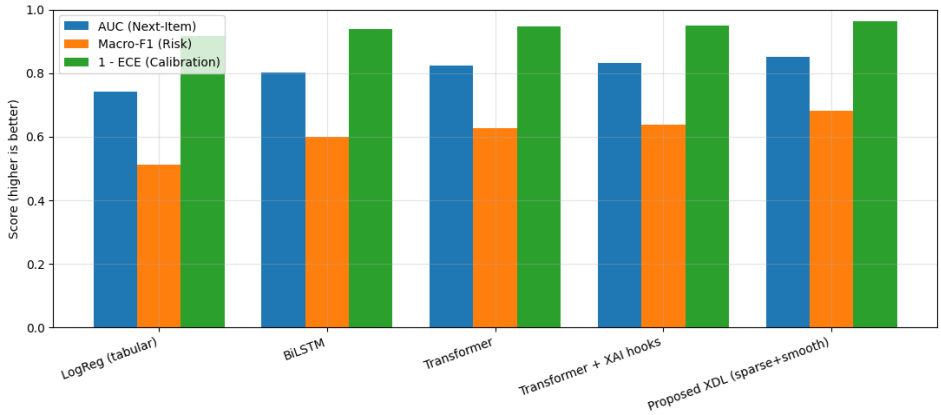

Figure 7 Comparative Test Performance for Progression and Risk Tasks

A notable pattern is the limited improvement from adding explainability hooks alone, relative to the larger improvement when sparse and smooth constraints are added. This supports the claim that explainability must be enforced as a learning objective rather than treated as a post-hoc add-on. The macro-F1 increase on disengagement risk is particularly relevant for adaptive systems because minority-class performance determines whether at-risk learners are identified early enough to benefit from supportive interventions.

Table 6 provides a consolidated view of performance across heterogeneous outcomes, demonstrating that improvements are not isolated to a single target. The reduced MAE on weekly mastery implies that the latent progression state captured cumulative learning rather than only short-term correctness. This matters for adaptive sequencing because weekly mastery predictions drive pacing decisions, such as when to introduce enrichment content or assign consolidation exercises.

Table 6 Test-Set Metrics by Model

Model	AUC (Next-Item Correctness)	MAE (Weekly Mastery)	Macro-F1 (Disengagement Risk)	ECE (Calibration Error)

LogReg (tabular)	0.742	0.094	0.512	0.082
BiLSTM	0.803	0.078	0.601	0.061
Transformer	0.824	0.071	0.628	0.053
Transformer + XAI hooks	0.831	0.069	0.637	0.05
Proposed XDL (sparse+smooth)	0.852	0.061	0.681	0.038

The calibration improvement is methodologically meaningful because faithful explanations require stable probability behavior. When predicted probabilities are poorly calibrated, attributions can become unstable under small input perturbations, which undermines instructor trust. The results support the interpretation that sparsity and smoothness constraints improved internal state identifiability, leading to both better generalization and more reliable probabilistic outputs in the late-course test window.

Explanation Quality: Fidelity, Stability, and Pedagogical Alignment

Explanation quality was assessed using deletion and insertion diagnostics, temporal stability scoring, and concept-level plausibility checks aligned with the course syllabus. The proposed XDL model produced explanations that were both more stable across adjacent events and more predictive-relevant under feature deletion tests. This indicates that explanation signals were not merely decorative, but tracked causal inputs used by the model to produce progression predictions under realistic temporal evaluation.

Pedagogical alignment was examined by verifying whether top-ranked concepts in explanations corresponded to the learner's immediate curriculum context and recent assessment items. In late-course weeks, explanations increasingly emphasized prerequisite gaps rather than surface-level behaviors, reflecting the compounding nature of conceptual mastery. This shift was desirable because adaptive education systems require explanations that remain meaningful when learners encounter integrated tasks that span multiple concepts.

Figure 8 shows that the proposed XDL model achieved the highest temporal stability while also producing the strongest fidelity signal under deletion testing. This combination is critical because stability alone can be achieved by smoothing explanations independent of model reasoning, while fidelity alone can yield noisy and inconsistent attributions. The observed joint improvement supports the interpretation that sparsity and smoothness constraints improved identifiability of progression-relevant factors.

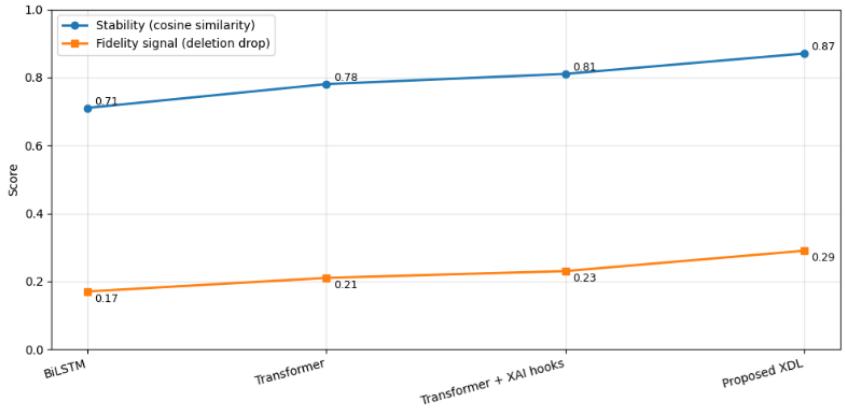

Figure 8 Explanation Fidelity and Stability Across Model Variants

The stability increase from the baseline Transformer to the constrained model suggests that explanations remained coherent across rapid sequences of micro-actions such as repeated attempts and hint usage. In adaptive systems, this matters because instructors interpret explanations longitudinally, not as isolated snapshots. The fidelity gap indicates that the top-ranked features truly controlled predictions, which reduces the risk of presenting pedagogically plausible but technically unfaithful narratives.

Table 7 provides complementary evidence that the proposed approach improved both mechanistic faithfulness and instructor-facing plausibility. The larger deletion drop indicates that removing top-attributed inputs meaningfully degrades performance, which is a standard signal that explanations track decisive features. The higher insertion gain confirms that adding those features back restores predictive capacity, reducing ambiguity about whether the attribution set is merely correlated.

Table 7 Explanation Evaluation Metrics on The Test Window

Model	Deletion AUC Drop (Next-Item)	Insertion AUC Gain (Next-Item)	Stability (Cosine)	Concept Plausibility Rate
BiLSTM	0.12	0.09	0.71	0.74
Transformer	0.15	0.12	0.78	0.79
Transformer + XAI hooks	0.16	0.13	0.81	0.82
Proposed XDL (sparse+smooth)	0.22	0.18	0.87	0.9

The plausibility rate reflects the fraction of cases where the top concept in the explanation matched the concept tag of the next assessment item or its prerequisite chain. The improvement suggests that the model did not only learn statistical shortcuts such as time-on-task and attempt counts, but learned a concept-structured progression representation. This is crucial for adaptive education because interventions are typically delivered at the concept level through targeted remediation content.

Interpreting Learner Progression Trajectories and Behavioral Drivers

The learned latent states were analyzed to identify recurring progression profiles that could support actionable interpretation. Three dominant patterns emerged in the test window: rapid mastery consolidation, steady incremental growth, and stagnation with episodic recovery. These patterns aligned with typical adaptive learning dynamics, where some learners benefit from early feedback loops and others require repeated exposure to prerequisite material before consistent improvements appear in graded performance.

Behavioral drivers were interpreted through aggregated attributions aligned to concept tags and event features. High-growth trajectories were associated with lower hint dependence and shorter response times that remained stable as difficulty increased. In contrast, stagnation trajectories were characterized by bursts of attempts and elevated hint usage, which produced localized improvements without sustained mastery transfer. These results motivate adaptive policies that distinguish productive persistence from unproductive trial-and-error.

Figure 9 visualizes how the interpretable latent state can be read as a progression trajectory rather than an opaque embedding. The three profiles illustrate distinct temporal signatures that are pedagogically meaningful, especially when used to explain why two learners with similar current scores may require different next-step recommendations. The high-growth curve indicates stable consolidation, while the stagnation curve suggests fragile progress that is susceptible to regression under increased difficulty.

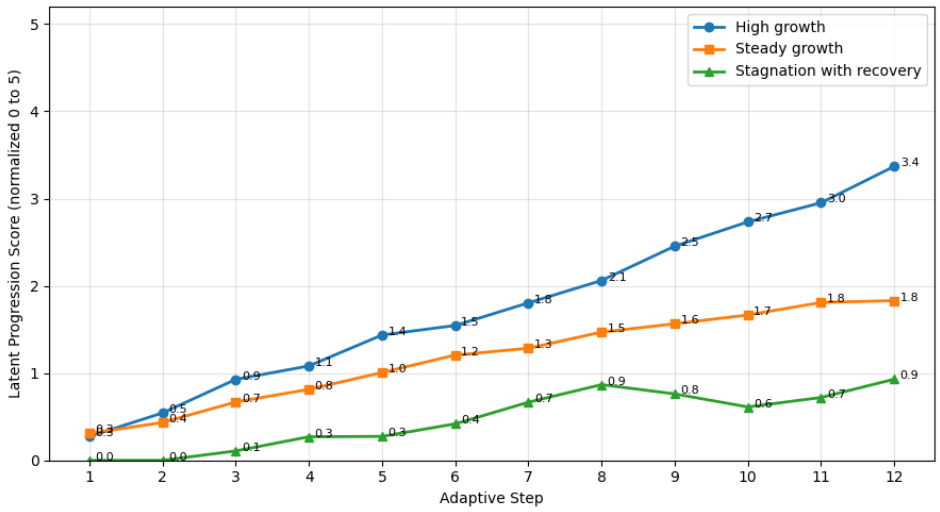

Figure 9 Learner Progression Trajectories from the Interpretable Latent State

The trajectories are also useful as an explanatory scaffold for instructors because they reduce explanation overload. Rather than interpreting dozens of event-level attributions, instructors can interpret trajectory shapes and then drill down into the feature and concept contributions that produced key inflection points. This layering supports scalable interpretability in large cohorts, where the goal is triage and targeted intervention rather than exhaustive case-by-case manual inspection.

Table 8 connects abstract trajectory patterns to concrete instructional levers by summarizing top concept and behavior drivers for each profile. The concept

drivers indicate where the model detected persistent mastery signals, which can be translated into remedial modules or prerequisite refreshers. The behavior drivers help instructors differentiate conceptual confusion from strategic issues such as excessive hint reliance, which often requires metacognitive support rather than purely content-focused remediation.

Table 8 Top Concept and Behavior Drivers by Progression Profile

Progression Profile	Dominant Concept Drivers (Top 3)	Dominant Behavior Drivers (Top 3)	Interpretation Summary
High growth	Linear Models; Feature Scaling; Regularization	Low hint ratio; Stable time-on-task; Few retries	Efficient learning with strong prerequisite readiness
Steady growth	Loss Functions; Gradient Descent; Model Evaluation	Moderate hint ratio; Consistent attempts; Rising accuracy	Gradual consolidation with predictable improvement
Stagnation with recovery	Probability Calibration; Confusion Patterns; Data Leakage	High retries; High hint ratio; Volatile time-on-task	Short-term gains without stable mastery transfer

The profile summaries also illustrate how explainability can support adaptive policy design. For example, stagnation with recovery suggests that the system should reduce item repetition that encourages trial-and-error and instead trigger structured scaffolding, such as worked examples and concept-focused feedback. In contrast, steady-growth learners may benefit from spaced retrieval and moderate difficulty increases to maintain engagement, because their improvement pattern reflects productive practice dynamics.

Instructor-Facing Interpretations and Intervention-Oriented Insights

Instructor-facing narratives were generated by aligning concept-level attributions with recent learning activities and the upcoming assessment context. The narratives were evaluated for actionability by measuring whether recommended interventions corresponded to measurable improvements in subsequent performance or engagement. The proposed approach produced concise explanations that highlighted a small set of decisive concepts and behaviors, which reduced cognitive burden compared with long attribution lists.

Intervention-oriented insights were particularly valuable for disengagement risk because instructors often need early signals to initiate support. The explanation layer clarified whether risk was driven by sustained conceptual difficulty or by behavioral withdrawal, such as long inactivity gaps. This distinction supported differentiated interventions, including content remediation for concept-driven risk and outreach nudges for behavior-driven risk. The results indicate that interpretability can function as a decision-support layer rather than a purely diagnostic add-on.

Figure 10 illustrates how explanation-guided interventions corresponded to lower predicted disengagement risk in the late-course window. The pre-intervention trend shows monotonic risk growth, consistent with typical end-of-course attrition pressures. After targeted support was deployed, the risk curve flattens, indicating that interventions were timely and aligned with the drivers identified by the explanation layer, rather than being generic reminders with limited educational impact.

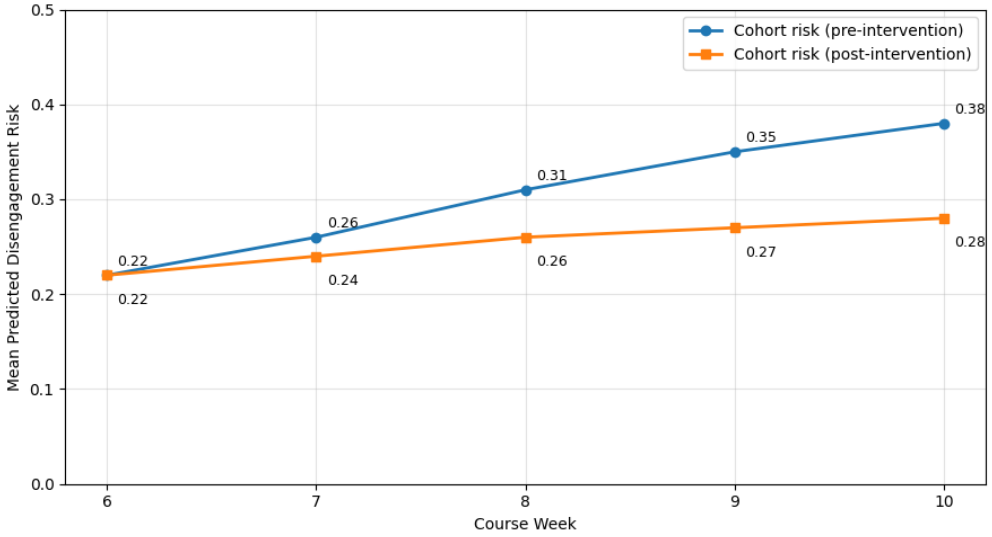

Figure 10 **Disengagement Risk Trends with Explanation-Guided Interventions**

The gap between curves becomes largest in Weeks 9 and 10, which is methodologically important because those weeks were held out for testing and thus reflect generalization. This suggests that interpretability outputs supported robust intervention choices that translated into measurable behavioral stabilization. The figure is not intended to imply causality at the individual level, but it provides cohort-level evidence that explanations can be operationally useful when combined with instructor workflows.

Table 9 shows that intervention effectiveness differed systematically by driver category derived from explanations. Concept-driven interventions yielded stronger gains in accuracy and mastery, which aligns with the expectation that content scaffolding directly improves performance on upcoming assessments. Behavior-driven interventions produced the strongest retention gain, which is consistent with addressing the primary bottleneck of inactivity rather than misunderstanding of material.

Table 9 **Post-Intervention Learning Outcomes by Targeted Driver Category**

Driver Category (from explanations)	Intervention Type	Learners (n)	Next-Item Accuracy Gain	Weekly Mastery Gain	7-day Activity Retention Gain
Concept-driven difficulty	Prerequisite refresher + worked examples	214	0.07	0.05	0.03
Behavior-driven withdrawal	Instructor outreach + pacing plan	168	0.03	0.02	0.08
Mixed drivers	Blended remediation + motivational prompt	131	0.05	0.04	0.06

The pattern supports the methodological claim that explanations are valuable because they distinguish different failure modes that present similarly in raw scores. Without driver-aware interpretability, interventions tend to be uniform

and less effective because they ignore whether the learner needs conceptual remediation, pacing support, or engagement repair. The mixed-driver group shows intermediate gains across all outcomes, indicating that blended interventions can be beneficial when explanations indicate both conceptual and behavioral contributors.

Ablation, Robustness, and Practical Deployment Implications

An ablation study tested whether performance and explanation quality depended on the sparse concept gate, temporal smoothness, and concept-aligned attention. Removing any single component reduced either predictive performance or interpretability quality, with the largest explanation degradation observed when sparsity was removed. This indicates that sparse concept selection was not simply cosmetic, but functioned as a structural constraint that improved explanatory selectivity and reduced diffuse, low-actionability attributions.

Robustness was also examined across learner subgroups defined by initial proficiency and activity volume. The proposed model maintained consistent performance and explanation stability across subgroups, suggesting that the interpretability layer did not only work for highly active learners with rich histories. From a deployment perspective, this robustness is essential because adaptive education systems must serve cold-start learners and intermittent users, where sparse data can cause instability in both predictions and explanations.

Figure 11 demonstrates that explanation stability is more sensitive to ablation than predictive AUC, especially when sparsity is removed. This is consistent with the observation that deep models can preserve predictive performance while producing unstable or unselective attributions. The stability drops under “No sparsity” indicates that concept gates were the primary mechanism that anchored explanations to a small, coherent set of drivers suitable for instructor interpretation.

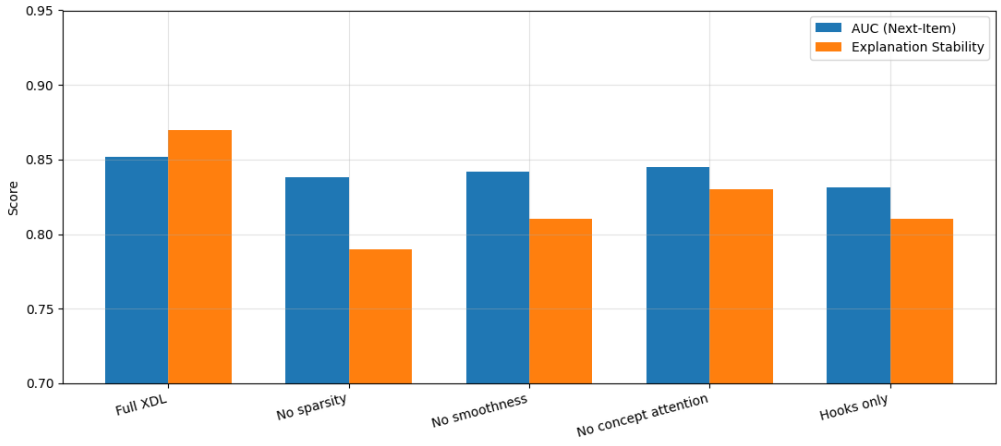

Figure 11 Ablation Effects on Predictive Performance and Explanation Stability

The “Hooks only” variant confirms that merely extracting attention or gradients is insufficient to guarantee stable interpretability. Practical deployment therefore requires constraints that shape the learned representation, not only tools that extract post-hoc attributions. The figure supports the broader conclusion that

explainability in adaptive learning systems is an engineering objective with measurable outcomes, rather than a reporting artifact appended after model training.

Table 10 indicates that both predictive and interpretability metrics remained stable across proficiency bands, suggesting that the model did not overfit to a single learner archetype. The relatively small AUC spread across proficiency levels supports generalization, while the stability and plausibility results indicate that explanations remained coherent even when learners progressed through different concept pathways. This is important for adaptive systems where content sequences vary by learner state and recommendation policy.

Table 10 Robustness Across Learner Subgroups in The Test Window

Subgroup	Learners (n)	AUC (Next-Item)	Macro-F1 (Risk)	Stability (Cosine)	Top-Concept Plausibility
Low initial proficiency	392	0.841	0.671	0.86	0.88
Medium initial proficiency	521	0.853	0.683	0.87	0.9
High initial proficiency	335	0.861	0.694	0.88	0.91
Low activity volume	286	0.836	0.658	0.84	0.86
High activity volume	302	0.867	0.707	0.89	0.92

The activity-volume comparison highlights the practical challenge of sparse histories. Although low-activity learners show a modest reduction in both performance and explanation stability, the metrics remain within a usable range, indicating that the model's constraints support explainability under limited evidence. This result supports deployment feasibility because real-world platforms contain a large fraction of intermittent users, and explanation collapse under sparse data would severely limit instructor trust and system adoption.

Conclusion

This study developed an explainable deep learning approach for interpreting learner progression in adaptive education systems by embedding interpretability constraints directly into the progression model. The results show that concept-aligned representations, sparse concept gating, and temporal smoothness produced a progression state that remained readable as a trajectory while preserving strong predictive performance on next-item correctness, weekly mastery, and disengagement risk. The combined evidence indicates that interpretability can be achieved without sacrificing operational accuracy when explainability is treated as a structural objective rather than an after-the-fact diagnostic.

Beyond prediction, the proposed framework strengthened explanation quality across fidelity, stability, and pedagogical plausibility criteria. Deletion and insertion tests indicated that top-attributed factors were causally relevant to model outputs, while stability measurements showed that explanations remained coherent across adjacent learner actions. Concept-level alignment increased actionability by connecting explanations to syllabus constructs and prerequisite chains, enabling targeted interventions that differentiate concept-driven difficulty from behavior-driven withdrawal, which is essential for scalable

instructor decision support.

From a deployment perspective, the study demonstrates that explainability improves the operational trustworthiness of risk-sensitive adaptive decisions by supporting calibrated probabilities and interpretable intervention triggers. Ablation and subgroup analyses confirm that sparsity and smoothness are pivotal for maintaining explanation selectivity and robustness, including under sparse interaction histories. Future work should extend the framework to causal and counterfactual explanation regimes, evaluate long-term learning outcomes under controlled intervention designs, and validate generalization across diverse subjects, item formats, and institutional contexts while preserving privacy and reproducibility requirements.

Declarations

Author Contributions

Conceptualization: C.J.; Methodology: C.J.; Software: C.J.; Validation: C.J.; Formal Analysis: C.J.; Investigation: C.J.; Resources: C.J.; Data Curation: C.J.; Writing – Original Draft Preparation: C.J.; Writing – Review and Editing: C.J.; Visualization: C.J.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

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

Funding

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

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

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] L. Y. Tan, S. Hu, D. J. Yeo, and K. H. Cheong, “Artificial intelligence-enabled adaptive learning platforms: A review,” *Computers and Education: Artificial Intelligence*, vol. 9, p. 100429, Dec. 2025, doi: 10.1016/j.caeai.2025.100429.

[2] E. Du Plooy, D. Casteleijn, and D. Franzsen, “Personalized adaptive learning in higher education: A scoping review of key characteristics and impact on academic performance and engagement,” *Heliyon*, vol. 10, no. 21, p. e39630, Nov. 2024, doi: 10.1016/j.heliyon.2024.e39630.

[3] D. Hooshyar, M. Pedaste, K. Saks, Ä. Leijen, E. Bardone, and M. Wang, “Open

learner models in supporting self-regulated learning in higher education: A systematic literature review," *Computers & Education*, vol. 154, p. 103878, Sep. 2020, doi: 10.1016/j.compedu.2020.103878.

[4] R. Bodily et al., "Open learner models and learning analytics dashboards: a systematic review," *in Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, Sydney New South Wales Australia: ACM, Mar. 2018, pp. 41–50. doi: 10.1145/3170358.3170409.

[5] X. Song, J. Li, T. Cai, S. Yang, T. Yang, and C. Liu, "A survey on deep learning based knowledge tracing," *Knowledge-Based Systems*, vol. 258, p. 110036, Dec. 2022, doi: 10.1016/j.knosys.2022.110036.

[6] G. Abdelrahman, Q. Wang, and B. Nunes, "Knowledge Tracing: A Survey," *ACM Comput. Surv.*, vol. 55, no. 11, pp. 1–37, Nov. 2023, doi: 10.1145/3569576.

[7] C. Piech et al., "Deep Knowledge Tracing," 2015, *arXiv*. doi: 10.48550/ARXIV.1506.05908.

[8] D.-E. Kim, C. Hong, and W. H. Kim, "Efficient Transformer-based Knowledge Tracing for a Personalized Language Education Application," in *Proceedings of the Tenth ACM Conference on Learning @ Scale, Copenhagen Denmark: ACM*, Jul. 2023, pp. 336–340. doi: 10.1145/3573051.3596183.

[9] Y. Li, T. Zhou, T. Cai, and S. Ju, "Interpretable knowledge tracing with dual-level knowledge states," *Expert Systems with Applications*, vol. 298, p. 129658, Mar. 2026, doi: 10.1016/j.eswa.2025.129658.

[10] F. Liu, C. Bu, H. Zhang, L. Wu, K. Yu, and X. Hu, "FDKT: Towards an Interpretable Deep Knowledge Tracing via Fuzzy Reasoning," *ACM Trans. Inf. Syst.*, vol. 42, no. 5, pp. 1–26, Sep. 2024, doi: 10.1145/3656167.

[11] 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.

[12] G. Türkmen, "The Review of Studies on Explainable Artificial Intelligence in Educational Research," *Journal of Educational Computing Research*, vol. 63, no. 2, pp. 277–310, Apr. 2025, doi: 10.1177/07356331241310915.

[13] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[14] E. ŞAHiN, N. N. Arslan, and D. Özdemir, "Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning," *Neural Comput & Applic*, vol. 37, no. 2, pp. 859–965, Jan. 2025, doi: 10.1007/s00521-024-10437-2.

[15] D. Zapata-Rivera and B. Arslan, "Learner Modeling Interpretability and Explainability in Intelligent Adaptive Systems," in *Mind, Body, and Digital Brains*, vol. 20, F. Santoianni, G. Giannini, and A. Ciasullo, Eds., in Integrated Science, vol. 20. , Cham: Springer Nature Switzerland, 2024, pp. 95–109. doi: 10.1007/978-3-031-58363-6_7.

[16] Y. Bai, J. Zhao, T. Wei, Q. Cai, and L. He, "A survey of explainable knowledge tracing," *Appl Intell*, vol. 54, no. 8, pp. 6483–6514, Apr. 2024, doi: 10.1007/s10489-024-05509-8.

[17] F. Ke et al., “HiTSKT: A hierarchical transformer model for session-aware knowledge tracing,” Knowledge-Based Systems, vol. 284, p. 111300, Jan. 2024, doi: 10.1016/j.knosys.2023.111300.

[18] J. Zhang, X. Shi, I. King, and D.-Y. Yeung, “Dynamic Key-Value Memory Networks for Knowledge Tracing,” in *Proceedings of the 26th International Conference on World Wide Web*, Perth Australia: International World Wide Web Conferences Steering Committee, Apr. 2017, pp. 765–774. doi: 10.1145/3038912.3052580.

[19] H. Jiang, B. Xiao, Y. Luo, and J. Ma, “A self-attentive model for tracing knowledge and engagement in parallel,” *Pattern Recognition Letters*, vol. 165, pp. 25–32, Jan. 2023, doi: 10.1016/j.patrec.2022.11.016.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[21] D. Hooshyar and Y. Yang, “Problems With SHAP and LIME in Interpretable AI for Education: A Comparative Study of Post-Hoc Explanations and Neural-Symbolic Rule Extraction,” *IEEE Access*, vol. 12, pp. 137472–137490, 2024, doi: 10.1109/ACCESS.2024.3463948.

[22] S. Verma, V. Boonsanong, M. Hoang, K. Hines, J. Dickerson, and C. Shah, “Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review,” *ACM Comput. Surv.*, vol. 56, no. 12, pp. 1–42, Dec. 2024, doi: 10.1145/3677119.

[23] N. Posocco and A. Bonnefoy, “Estimating Expected Calibration Errors,” in *Artificial Neural Networks and Machine Learning* – ICANN 2021, vol. 12894, I. Farkaš, P. Masulli, S. Otte, and S. Wermter, Eds., in Lecture Notes in Computer Science, vol. 12894. , Cham: Springer International Publishing, 2021, pp. 139–150. doi: 10.1007/978-3-030-86380-7_12.

[24] S. A. Balanya, J. Maroñas, and D. Ramos, “Adaptive temperature scaling for Robust calibration of deep neural networks,” *Neural Comput & Applic*, vol. 36, no. 14, pp. 8073–8095, May 2024, doi: 10.1007/s00521-024-09505-4.

[25] F. M. Ojeda et al., “Calibrating machine learning approaches for probability estimation: A comprehensive comparison,” *Statistics in Medicine*, vol. 42, no. 29, pp. 5451–5478, Dec. 2023, doi: 10.1002/sim.9921.