IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, April 2025
Impact Factor: 7.67

An Explainable and Personalized Generative AI Framework for Adaptive Learning, Intelligent Tutoring, and Student Performance Prediction

Dr. Sonam Bansal
Associate Professor, Department of Education
Gurugram University, Gurugram, Haryana, India
sonambansal@gurugramuniversity.ac.in

Abstract: *Generative artificial intelligence has made it technically feasible to deliver individualised instruction at scale, yet three problems continue to limit its adoption in formal education. First, most deployed systems are generic: they respond to the prompt in front of them without any durable model of who the learner is, what the learner already knows, or where the learner is likely to fail next. Second, they are opaque: neither the learner nor the teacher can see why a particular hint, difficulty level, or risk score was produced, which corrodes trust and makes pedagogical accountability impossible. Third, predictive analytics and tutoring are usually built as separate systems, so a model that flags a student as at risk rarely does anything about it. This paper proposes EPGAI (Explainable and Personalised Generative AI), an integrated framework that unifies learner modelling, generative tutoring, and student performance prediction inside a single explainability-first architecture. EPGAI comprises five interacting layers: a multi-modal learner model that maintains cognitive, behavioural, and affective state; an attention-based knowledge tracing and performance prediction module that estimates mastery and forecasts assessment outcomes; a retrieval-augmented generative tutoring engine constrained by curriculum documents and pedagogical prompt templates; an explanation layer that converts SHAP attributions, counterfactual analysis, and rationale traces into classroom-appropriate natural language for three distinct audiences; and an adaptation policy that closes the loop by translating predictions and explanations into concrete instructional actions. The framework is evaluated on three public educational datasets (OULAD, ASSISTments, and EdNet) together with a mixed-methods study involving undergraduate learners and teachers. Results indicate that the integrated model improves predictive accuracy over conventional baselines, that retrieval grounding substantially reduces factually incorrect tutoring content, and that learner-facing explanations improve reported trust and self-regulation without harming learning outcomes. The paper also examines fairness across demographic subgroups, data-protection obligations, and the pedagogical risks of over-reliance on generative tutors, and concludes with a research agenda for classroom-scale validation.*

Keywords: generative artificial intelligence; explainable AI; adaptive learning; intelligent tutoring systems; knowledge tracing; student performance prediction; learning analytics; educational data mining; retrieval-augmented generation; algorithmic fairness in education

I. INTRODUCTION

The aspiration to give every learner the equivalent of a skilled personal tutor is one of the oldest goals of educational research. Bloom's widely cited two-sigma finding reported that students taught individually with mastery-based feedback outperformed conventionally instructed peers by roughly two standard deviations [1]. The finding has been contested and refined in the four decades since, but its practical implication has never been seriously disputed: individual attention works, and it does not scale within the economics of public education. Intelligent tutoring systems were the field's first systematic attempt to close that gap computationally. Meta-analytic evidence suggests that well-

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 3, April 2025 Impact Factor: 7.67

designed tutoring systems can approach, though generally not equal, the effectiveness of human tutors [2]. Their limitation has always been brittleness. Classical systems depend on hand-authored domain models, production rules, and hint sequences, each of which must be written by a subject expert for every topic, misconception, and problem type. The authoring cost is prohibitive for most institutions, and the resulting dialogue is confined to paths the author anticipated in advance.

Large language models have changed this calculus. Transformer architectures [9] trained at scale and aligned with human feedback [10, 11] can generate explanations, worked examples, Socratic questions, analogies, and formative feedback across virtually any curricular domain without topic-specific authoring. Educational researchers have consequently identified generative AI as a genuine inflection point for personalised learning, while cautioning that opportunity and risk arrive together [14]. Three deficiencies, in particular, prevent current generative systems from functioning as trustworthy instructional agents in formal education.

The memorylessness problem. A general-purpose conversational assistant treats each request as an isolated event. It has no persistent representation of the learner's mastery profile, error history, pace, preferred representational modality, or affective state. It therefore cannot distinguish a student who has never encountered a concept from one who has failed at it three times in a fortnight, and it cannot deliberately sequence instruction toward a long-term learning objective. Personalisation reduced to tone and reading level is not pedagogical personalisation; it is stylistic accommodation.

The opacity problem. Deep knowledge tracing models [5–8] and neural predictors of course outcome are accurate but uninterpretable. When such a model assigns a student a seventy-one per cent probability of failing an assessment, neither the student, the teacher, nor the institution can inspect the reasoning behind the number. Generative tutors compound the difficulty, because fluent prose is persuasive irrespective of its correctness, and models are known to produce confident fabrications [15]. In a domain where a wrong explanation can install a durable misconception, unexamined fluency is a hazard rather than a feature. Interpretability is therefore not a desirable extra in educational AI; researchers have argued for some years that it is a precondition of responsible deployment [18–20].

The disconnection problem. Learning analytics and instructional delivery are typically implemented as separate systems owned by separate teams. Early-warning dashboards identify at-risk students and pass a list to an academic adviser; the tutoring platform, if one exists, knows nothing of that list. Prediction without an intervention pathway produces institutional anxiety rather than institutional benefit, and it places the entire remediation burden on already overloaded staff.

This paper proposes and evaluates EPGAI, an Explainable and Personalised Generative AI framework designed to address the three deficiencies jointly rather than sequentially. The central design commitment is that prediction, generation, and explanation must share a single learner representation. A forecast of assessment difficulty is useful only if it can immediately shape what the tutor says next; a generated hint is defensible only if its provenance and pedagogical rationale can be inspected; and an explanation is meaningful only if it is expressed in terms the intended audience can act upon. Accordingly, EPGAI produces three parallel explanation registers from the same underlying evidence: a metacognitive register for learners, a diagnostic register for teachers, and a technical register for auditors and researchers.

The specific contributions of this work are as follows.

A five-layer reference architecture that integrates multi-modal learner modelling, attention-based knowledge tracing, retrieval-grounded generative tutoring, multi-audience explanation, and a closed-loop adaptation policy within one system.

A hybrid performance prediction module that fuses sequential interaction embeddings with engineered behavioural and engagement features, improving over both purely sequential and purely tabular baselines while remaining amenable to attribution analysis.

An explanation pipeline that converts SHAP attributions and counterfactual queries into audience-specific natural language, with a verification stage that constrains the language model to the numerical evidence it is given.

A retrieval-augmented, pedagogically templated tutoring engine whose responses are constrained by institutional curriculum documents, reducing unsupported content relative to unconstrained generation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, April 2025
Impact Factor: 7.67

An empirical evaluation across three public datasets and a mixed-methods user study, accompanied by a subgroup fairness analysis and an explicit treatment of ethical, privacy, and over-reliance risks.

The remainder of the paper is organised as follows. Section 2 reviews related work in adaptive learning, knowledge tracing, generative AI in education, and explainable learning analytics, and identifies the research gap. Section 3 presents the EPGAI architecture in detail. Section 4 describes datasets, model configurations, baselines, and evaluation protocol. Section 5 reports results. Section 6 discusses implications for teaching practice, Section 7 addresses ethical and governance considerations, Section 8 states limitations, and Section 9 concludes.

II. RELATED WORK

2.1 Adaptive Learning and Intelligent Tutoring Systems

Adaptive learning research is grounded in the constructivist premise that instruction is most effective when pitched just beyond current independent competence, the region Vygotsky termed the zone of proximal development [3]. Operationalising this idea computationally requires a continuously updated estimate of what the learner can and cannot do. Early intelligent tutoring systems achieved this through model tracing against expert production rules, an approach that produced impressive results in well-structured domains such as algebra and programming but generalised poorly to ill-structured subjects. VanLehn's synthesis of tutoring effectiveness concluded that granularity of interaction, rather than surface sophistication, best predicts learning gain: systems that intervene at the level of individual reasoning steps outperform those that respond only to final answers [2]. Cognitive load theory [27] and the ICAP framework [26] supply complementary design constraints, indicating respectively that extraneous processing demands should be minimised and that constructive and interactive engagement produce deeper learning than passive reception. EPGAI adopts all three commitments: step-level interaction, load-aware presentation, and prompts that require the learner to generate rather than merely receive.

2.2 Knowledge Tracing and Student Performance Prediction

Bayesian Knowledge Tracing modelled mastery as a latent binary variable updated by a hidden Markov process with fixed learning, guessing, and slipping parameters [4]. Its transparency remains a benchmark, but its assumptions of skill independence and monotonic learning are restrictive. Deep Knowledge Tracing replaced the explicit latent variable with a recurrent hidden state and produced substantial accuracy gains [5]. Subsequent architectures introduced external memory for skill-level representation [6], self-attention over interaction sequences [7], and context-aware attention with monotonic decay reflecting forgetting [8]. Parallel to this sequential literature, the educational data mining community has predicted course-level outcomes from clickstream, submission, and demographic features using gradient-boosted trees and similar tabular learners [28], typically achieving strong performance on institutional data such as OULAD [21]. The two traditions rarely meet: sequential models capture fine-grained skill dynamics but ignore engagement context, while tabular models capture engagement but discard temporal structure. EPGAI fuses them explicitly.

2.3 Generative AI and Retrieval Grounding in Education

Instruction-tuned language models can produce pedagogically plausible explanations, distractor sets, worked examples, and formative feedback with minimal task-specific supervision [10, 11]. Chain-of-thought prompting improves multi-step reasoning quality and, incidentally, exposes intermediate steps that can be inspected pedagogically [13]. The principal reliability concern is hallucination: models generate fluent statements unsupported by any source [15], a failure mode with unusually high cost in education because learners lack the domain knowledge required to detect it. Retrieval-augmented generation mitigates this by conditioning generation on retrieved documents [12], and in an educational deployment the retrieval corpus can be restricted to the institution's own syllabus, prescribed textbooks, and validated question banks. Kasneci and colleagues survey both the opportunities and the governance requirements of large models in education, emphasising teacher oversight and competence development [14]. EPGAI treats retrieval grounding and pedagogical templating as mandatory rather than optional components of the generation path.

2.4 Explainable AI in Learning Analytics

Post-hoc attribution methods dominate applied explainability. LIME fits an interpretable surrogate in the neighbourhood of a prediction [17], while SHAP derives feature attributions with the additivity and consistency properties of Shapley values [16]. Counterfactual explanation offers a complementary and often more actionable form:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, April 2025
Impact Factor: 7.67

instead of decomposing a prediction, it identifies the minimal change in input that would alter the outcome [29]. Doshi-Velez and Kim caution that interpretability is not a single measurable property and must be defined relative to a stakeholder and a task [18]. Within education specifically, Conati and colleagues argue that interpretable models are necessary for both learner trust and teacher agency [19], and Khosravi and colleagues provide a systematic framework for explainable AI in education organised around stakeholders, purposes, and presentation formats [20]. The recurrent gap in this literature is translation: attribution vectors are produced for data scientists, not for sixteen-year-olds or subject teachers. EPGAI addresses this by treating explanation generation as a distinct, verified natural-language task with three audience-specific registers.

2.5 Research Gap

Table 1 positions representative strands of prior work against the three deficiencies identified in Section 1. Existing systems address personalisation, generation, or explanation, but seldom all three, and almost never with a shared learner representation that permits predictions to influence tutoring within the same session.

Table 1. Positioning of representative research strands against the capabilities required for trustworthy adaptive instruction.

Research strand	Primary objective	Persistent learner model	Generative content	Stakeholder explanation	Closed loop
Classical ITS (model tracing)	Step-level remediation	Yes	No	Partial	Yes
BKT [4]	Mastery estimation	Yes	No	Yes	Partial
DKT / SAKT / AKT [5, 7, 8]	Sequence prediction	Implicit	No	No	No
Tabular early warning [21, 28]	Risk flagging	Partial	No	Partial	No
General LLM assistants [10, 14]	On-demand help	No	Yes	No	No
XAI in education [19, 20]	Transparency	Varies	No	Yes	Partial
EPGAI (this work)	Integrated instruction	Yes	Yes	Yes (3 registers)	Yes

III. THE PROPOSED EPGAI FRAMEWORK

EPGAI is organised as five layers arranged in a continuous cycle: the learner model (L1) supplies state to the prediction engine (L2); predictions and learner state jointly condition the generative tutoring engine (L3); every prediction and generated artefact passes through the explanation layer (L4); and the adaptation policy (L5) converts the resulting evidence into the next instructional action, whose outcome updates L1. The cycle is deliberately closed: no component produces output that terminates outside the system.

3.1 Layer 1: Multi-Modal Learner Model

The learner model maintains a persistent state vector for learner u at time t, denoted S_u(t), composed of four sub-vectors. The cognitive sub-vector holds per-concept mastery estimates over the curriculum knowledge graph, together with an uncertainty term reflecting how much evidence supports each estimate. The behavioural sub-vector records interaction regularity, session frequency and duration, resource access breadth, submission timing relative to deadlines, and revision behaviour. The affective and motivational sub-vector holds proxy indicators of frustration, disengagement, and persistence, inferred from response latency patterns, abandonment, help-seeking frequency, and short self-report probes rather than from any biometric sensing. The preference sub-vector stores modality and scaffolding preferences, including representational format, worked-example dependence, and language of instruction.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 3, April 2025 Impact Factor: 7.67

The curriculum knowledge graph is authored once per course as a directed acyclic graph of concepts with prerequisite edges. Mastery propagates along these edges during inference, so that evidence of failure on a dependent concept raises uncertainty about its prerequisites. This structure is what allows the system to diagnose a surface error as a deeper prerequisite gap, which is the single most valuable diagnostic act a tutor performs.

3.2 Layer 2: Hybrid Performance Prediction Engine

The prediction engine addresses two related tasks. The first is next-step correctness, the standard knowledge tracing formulation: given the interaction sequence x_1 through x_t, estimate the probability that the learner answers the next item correctly. The second is horizon-level outcome prediction: estimate the probability of failure or withdrawal at the module or semester level, sufficiently early for intervention to be possible.

For the sequential component, interactions are encoded as tuples of item identifier, associated concepts, correctness, response latency, and elapsed time since the previous interaction. These are embedded and processed by a self-attention encoder with monotonic attention decay, following the context-aware attentive formulation [8], which encodes the intuition that recent evidence is more informative than distant evidence while still allowing older interactions to contribute. The resulting contextual representation h_t is a compact summary of the learner's trajectory.

For the tabular component, engineered features summarising engagement, regularity, assessment history, and resource use are computed over rolling windows and passed to a gradient-boosted tree ensemble [28]. The two components are fused by concatenating h_t with the tabular feature vector and passing the result through a shallow feed-forward network with dropout, trained with binary cross-entropy under class weighting to counteract the imbalance typical of failure prediction. Fusion at the representation level, rather than simple score averaging, allows interactions between sequence dynamics and engagement context to be learned directly; the ablation in Section 5.2 quantifies the benefit.

Table 2. Feature groups supplied to the prediction engine.

Feature group	Representative features	Rationale
Sequential interaction	Item and concept embeddings, correctness sequence, response latency, inter-attempt interval, attempt count	Captures skill acquisition and forgetting dynamics
Engagement	Sessions per week, mean session duration, distinct resources accessed, video completion ratio	Distinguishes sustained study from sporadic contact
Regularity	Variance of inter-session gaps, longest inactivity gap, proportion of late submissions	Regularity is a stronger predictor than raw volume
Assessment history	Rolling mean and trend of scores, first-attempt accuracy, resubmission rate	Trend often precedes level as a warning signal
Help-seeking	Hint requests per item, tutor turns per session, abandonment rate	Separates productive from avoidant struggle
Contextual	Course load, module difficulty index, prior qualification band	Controls for structural rather than individual factors

3.3 Layer 3: Retrieval-Grounded Generative Tutoring Engine

The tutoring engine constructs each response through a four-stage pipeline. In the first stage, intent classification assigns the learner turn to one of a small set of pedagogical intents: conceptual explanation, worked example, hint on a current attempt, error diagnosis, practice generation, or metacognitive support. In the second stage, retrieval collects the most relevant passages from an indexed corpus of the institution's syllabus, prescribed textbook sections, lecture notes, and validated solved examples [12]. In the third stage, a pedagogical prompt is assembled from a template selected by intent, populated with retrieved passages, the relevant slice of the learner state S_u(t), and explicit instructional constraints. In the fourth stage, a verification pass checks the draft response for claims unsupported by the retrieved passages, for premature disclosure of the final answer when the intent is hinting, and for reading level appropriate to the learner profile; failures trigger regeneration with a corrective instruction.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 3, April 2025 Impact Factor: 7.67

Prompt templates encode pedagogy explicitly rather than leaving it to model disposition. A hint template, for example, instructs the model to identify the first incorrect step, ask a question directed at the reasoning behind that step, withhold the corrected step, and offer a concrete representation only if the learner's recorded load indicators are elevated. This templating is what converts a general assistant into a tutor: it enforces the constructive engagement the ICAP framework associates with deeper learning [26] and prevents the model from defaulting to the most immediately satisfying behaviour, which is simply to supply the answer.

3.4 Layer 4: Explanation Layer

The explanation layer operates on every prediction and every generated artefact. For predictions, SHAP values are computed to attribute the model output across feature groups [16], with LIME used as a stability cross-check on a sampled subset [17]. Counterfactual queries are then generated by searching for the minimal feasible change in actionable features that would move the prediction across the decision threshold [29]; actionability is enforced by restricting the search to features the learner can influence, which excludes prior qualification and demographic attributes by construction. For generated tutoring content, the layer records the retrieved passages used, the template invoked, the learner-state slice supplied, and the verification outcome, producing an auditable provenance trace for every utterance.

These artefacts are then rendered in three registers. The learner register is metacognitive and forward-looking: it names one or two dominant factors and one concrete action, avoiding both numerical risk scores and language that invites fixed-ability attribution. The teacher register is diagnostic and comparative: it presents attribution rankings, the affected concept nodes in the knowledge graph, cohort context, and a suggested intervention. The auditor register is technical and complete: full attribution vectors, model version, data window, confidence intervals, and subgroup performance. The natural-language rendering is itself produced by a language model, but constrained to a structured evidence object and post-checked so that every quantitative claim in the generated text corresponds to a value in that object. This constraint is essential, since an unverified explanation generator can produce plausible narratives that misrepresent the model it purports to explain.

Table 3. The three explanation registers produced from a single prediction.

Register	Audience and purpose	Content	Illustrative form
Metacognitive	Learner; self-regulation	One or two dominant actionable factors, one next action, no risk score	Gaps between practice sessions have grown; two short sessions this week on quadratic factorisation would help most
Diagnostic	Teacher; intervention design	Ranked attributions, affected concept nodes, cohort comparison, suggested action	Risk 0.68. Main drivers: submission delay, declining first-attempt accuracy on Module 3 prerequisites
Technical	Auditor, researcher	Full SHAP vector, counterfactuals, model version, data window, subgroup metrics	Complete attribution table with confidence intervals and fairness diagnostics

3.5 Layer 5: Adaptation Policy

The adaptation policy selects the next instructional action from a bounded action space: adjust item difficulty, insert prerequisite remediation, change representational modality, alter scaffolding intensity, trigger spaced retrieval of a decaying concept, or escalate to human contact. Selection is governed by a utility function balancing expected mastery gain against estimated cognitive load and motivational cost, with difficulty targeted at a success probability band empirically associated with productive challenge. Two safeguards constrain the policy. First, escalation to a human teacher is mandatory, not discretionary, when sustained disengagement, repeated failure after remediation, or affective distress indicators are detected. Second, teachers may override any policy decision, and overrides are logged as training signal, preserving professional judgement as the final authority in the loop.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, April 2025
Impact Factor: 7.67

3.6 System Workflow

Algorithm 1 summarises one complete interaction cycle.

Algorithm 1: EPGAI interaction cycle

```
1.  Observe interaction x_t for learner u; append to sequence.
2.  Update learner state S_u(t): cognitive, behavioural, affective, preference.
3.  Propagate mastery uncertainty along prerequisite edges of the knowledge
graph.
4.  Compute p_next (next-step correctness) and p_risk (horizon outcome).
5.  Compute SHAP attributions and actionable counterfactuals for p_risk.
6.  Classify pedagogical intent of the learner turn.
7.  Retrieve grounding passages from the curriculum corpus.
8.  Assemble templated prompt with retrieved passages and S_u(t) slice.
9.  Generate response; verify grounding, hint discipline, and reading level.
10. Render explanations in learner, teacher, and auditor registers.
11. Select next action by utility; escalate to teacher if safeguard triggered.
12. Log provenance trace; return to step 1.
```

IV. METHODOLOGY

4.1 Datasets

Evaluation uses three publicly available educational datasets selected to span different granularities and institutional contexts. The Open University Learning Analytics Dataset (OULAD) provides demographic records, virtual learning environment clickstream data, assessment submissions, and final outcomes for distance-education modules, and is well suited to horizon-level outcome prediction [21]. The ASSISTments dataset provides fine-grained item-level responses within a mathematics tutoring platform and is a standard benchmark for knowledge tracing [22]. EdNet supplies a large-scale hierarchical record of learner interactions from a language-learning platform, permitting evaluation of scalability and of behaviour under long interaction histories [23]. Using three datasets of different structure guards against conclusions that are artefacts of a single platform's design.

Table 4. Datasets and prediction tasks.

Dataset	Domain	Granularity	Primary task	Split protocol
OULAD [21]	Distance HE modules	Session and assessment	Failure / withdrawal risk	Presentation-wise, temporal
ASSISTments [22]	School mathematics	Item level	Next-step correctness	Learner-wise 5-fold
EdNet [23]	Language learning	Item level, long histories	Next-step correctness, scalability	Learner-wise 5-fold

4.2 Preprocessing and Protocol

Learners with fewer than ten recorded interactions were excluded from sequence modelling as insufficiently informative, and sequences were truncated to a maximum length of two hundred interactions with sliding windows for longer histories. Continuous features were standardised using statistics computed on training folds only. Response latencies were log-transformed and winsorised at the first and ninety-ninth percentiles to limit the influence of idle sessions recorded as extreme durations. Splits are strictly learner-wise for the item-level datasets, so that no learner appears in both training and test partitions, and temporally forward for OULAD, so that models are never trained on data postdating the evaluation window. This second constraint matters more than it is usually given credit for: random splitting on longitudinal educational data permits leakage from future behaviour and inflates reported performance in ways that do not survive deployment.

Horizon-level risk prediction is evaluated at three checkpoints corresponding to approximately twenty-five, fifty, and seventy-five per cent of module elapsed time, since a model that is accurate only at the end of term is of no practical

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, April 2025
Impact Factor: 7.67

use. Class imbalance is addressed through class-weighted loss rather than synthetic oversampling, to avoid generating implausible learner profiles.

4.3 Baselines, Configuration, and Metrics

Baselines comprise logistic regression on engineered features, a gradient-boosted tree ensemble [28], Bayesian Knowledge Tracing [4], Deep Knowledge Tracing [5], a self-attentive tracer [7], and a context-aware attentive tracer [8]. The EPGAI prediction engine uses a four-layer attention encoder with eight heads, embedding dimension 128, dropout 0.2, Adam optimisation with an initial learning rate of 0.001 and cosine decay, batch size 64, and early stopping on validation area under the curve with patience of ten epochs. Hyperparameters were selected by random search on validation folds. The generative components use an instruction-tuned large language model accessed through a fixed system prompt and temperature 0.3 for tutoring and 0.1 for explanation rendering; retrieval uses dense embeddings with a hybrid lexical re-ranking stage over a corpus of curriculum documents.

Predictive performance is reported as area under the ROC curve, F1 score for the minority class, accuracy, and root mean squared error. Explanation quality is assessed for fidelity, using the drop in model output when top-attributed features are ablated; for stability, using rank correlation of attributions across bootstrap resamples; and for comprehensibility, using stakeholder ratings. Tutoring quality is rated by three subject experts on pedagogical appropriateness, factual correctness, and hint discipline (whether the response withheld the answer when it should have), with inter-rater agreement reported as Krippendorff’s alpha. Fairness is assessed as the maximum gap in true positive rate and in false positive rate across subgroups defined by prior qualification band, age band, and socio-economic index available in OULAD.

4.4 Human Study

A mixed-methods study was conducted with undergraduate participants studying an introductory quantitative methods module and with teaching staff associated with that module. Participants were assigned to three conditions: a control condition using standard course materials, a generative tutor condition without explanations, and the full EPGAI condition with explanation registers enabled. Learning was measured by pre-test and post-test on matched item sets; perceived trust, transparency, and self-regulation were measured by validated Likert instruments administered post-intervention; and semi-structured interviews with teaching staff were analysed thematically. Participation was voluntary and separate from course assessment, informed consent was obtained, data were pseudonymised at collection, and participants could withdraw without academic consequence.

V. RESULTS

5.1 Predictive Performance

Table 5 reports next-step correctness prediction on the item-level datasets and horizon-level risk prediction on OULAD at the fifty per cent checkpoint. The hybrid fusion model outperforms both families of baseline on all three datasets. The margin over the strongest sequential baseline is modest on ASSISTments, where engagement context is thin, and larger on OULAD, where behavioural regularity carries substantial signal that sequence models alone cannot access. This pattern is consistent with the motivating hypothesis: fusion helps most where the two information sources are genuinely complementary.

Table 5. Predictive performance across datasets and models (mean over folds).

Model	ASSISTments AUC	ASSISTments F1	EdNet AUC	EdNet F1	OULAD AUC	OULAD F1
Logistic regression	0.702	0.664	0.688	0.651	0.771	0.612
Gradient boosting [28]	0.741	0.706	0.724	0.690	0.842	0.703
BKT [4]	0.716	0.679	0.701	0.663	—	—
DKT [5]	0.803	0.752	0.776	0.731	0.821	0.674
SAKT [7]	0.818	0.769	0.795	0.748	0.833	0.688

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 3, April 2025 Impact Factor: 7.67

Model	ASSISTments AUC	ASSISTments F1	EdNet AUC	EdNet F1	OULAD AUC	OULAD F1
AKT [8]	0.827	0.778	0.806	0.759	0.845	0.706
EPGAI (fusion)	0.841	0.793	0.822	0.774	0.879	0.741

Note: values are reported for the configuration described in Section 4.3; OULAD figures correspond to the 50% elapsed-time checkpoint. BKT is not applicable to module-level outcome prediction.

Performance at earlier checkpoints is naturally lower but remains actionable. At the twenty-five per cent checkpoint the fusion model retains an area under the curve of approximately 0.81 on OULAD, which is sufficient to prioritise outreach while leaving most of the term available for intervention. Precision at the top decile of predicted risk is the more operationally relevant quantity for institutions with limited advising capacity, and this remains above 0.70 across checkpoints.

5.2 Ablation Study

Table 6 isolates the contribution of each architectural component on OULAD. Removing the tabular engagement branch produces the largest single degradation, confirming that behavioural regularity is not recoverable from item sequences alone. Removing knowledge-graph propagation degrades performance modestly but has a disproportionate effect on diagnostic quality, since without it the system identifies symptoms rather than prerequisite causes. Removing retrieval grounding leaves predictive metrics untouched, as expected, but sharply increases unsupported claims in tutoring output.

Table 6. Ablation of EPGAI components (OULAD, 50% checkpoint; tutoring metrics on held-out dialogue sample).

Configuration	AUC	F1	Unsupported claims (%)	Hint discipline (%)
Full EPGAI	0.879	0.741	4.1	91.3
− Tabular engagement branch	0.848	0.709	4.3	91.0
− Sequential attention branch	0.851	0.712	4.2	90.8
− Knowledge-graph propagation	0.866	0.727	5.0	88.4
− Retrieval grounding	0.878	0.740	17.6	86.1
− Pedagogical templating	0.879	0.741	9.2	58.7
− Explanation verification	0.879	0.741	11.4	90.6

The templating ablation is the most instructive result in the table. Without pedagogical templates the model answers correctly and helpfully in the conversational sense, but it supplies final answers instead of hints in a large proportion of cases where the learner was mid-attempt. An assistant that is maximally helpful turn-by-turn is not thereby a good tutor; productive difficulty must be engineered deliberately, and it will not emerge from a general helpfulness objective.

5.3 Explanation Quality

Attribution fidelity, measured as the mean drop in predicted risk when the top three attributed features are ablated, was substantially larger than for randomly selected feature triples, indicating that attributions identify genuinely load-bearing inputs. Stability, measured as Spearman correlation of attribution ranks across bootstrap resamples, was high for the fusion model's tabular features and moderate for sequence-derived features, whose attributions are inherently more diffuse. Comprehensibility ratings from learners and teachers are summarised in Table 7. Teachers rated the diagnostic register as materially more useful than a bare risk score for planning intervention, and several noted in interview that the value lay less in the number than in seeing which concept nodes were implicated.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, April 2025
Impact Factor: 7.67

Table 7. Stakeholder ratings of explanation quality (5-point scale, higher is better).

Dimension	Numeric score only	SHAP plot	EPGAI register	Change
Learner comprehension	2.4	2.1	4.2	+1.8
Learner perceived fairness	2.7	2.9	4.0	+1.3
Learner reported actionability	2.2	2.5	4.3	+2.1
Teacher diagnostic usefulness	2.9	3.6	4.4	+1.5
Teacher trust in system	2.8	3.4	4.1	+1.3

Note: "Change" is relative to the numeric-score-only condition. SHAP plots were presented in standard force-plot form without narrative translation.

The finding that raw SHAP plots scored below a bare numeric score on learner comprehension deserves emphasis. Technical explainability artefacts are not self-explanatory to non-technical audiences and can actively reduce comprehension by presenting unfamiliar visual encodings. Explanation is a communicative act, and it requires translation rather than exposure.

5.4 Learning Outcomes and Engagement

Normalised learning gain from pre-test to post-test was highest in the full EPGAI condition, followed by the generative tutor without explanations, with the control condition lowest. The gap between the two generative conditions was smaller than the gap between either and control, suggesting that the principal learning benefit derives from adaptive tutoring itself, while the explanation layer contributes primarily to trust, perceived fairness, and self-regulatory behaviour. This is a meaningful distinction: explanations should be justified on grounds of accountability and learner agency, not solely on grounds of test score improvement. Engagement measures showed higher session regularity and lower task abandonment in the EPGAI condition, and interview data indicated that learners in the explanation condition were more likely to describe difficulty in terms of specific gaps than in terms of general ability, which is the attributional pattern associated with persistence.

5.5 Fairness Analysis

Subgroup analysis on OULAD examined true and false positive rate gaps across prior qualification, age band, and socio-economic index. The baseline gradient-boosted model exhibited a noticeably higher false positive rate for learners from lower prior-qualification bands, meaning that such learners were disproportionately flagged as at risk when they in fact succeeded. Applying class-weighted training with a subgroup-aware threshold calibration reduced this gap without materially reducing overall performance. The residual gap is not eliminated, and the paper does not claim that it can be eliminated by technical means alone: differential false positive rates frequently reflect real differences in the historical support available to different groups, encoded in the outcome labels themselves. Baker and Hawn's analysis of algorithmic bias in education is directly relevant here, particularly their observation that unmeasured contextual variables often drive apparent model unfairness [24]. Operationally, EPGAI mitigates the consequence of false positives by ensuring that a risk flag triggers an offer of support rather than any restriction of opportunity, and by excluding demographic attributes from counterfactual explanations so that no learner is ever told that an immutable characteristic explains their prediction.

VI. DISCUSSION

6.1 Implications for Teaching Practice

The results suggest a division of labour rather than a substitution. The system is well suited to work that is repetitive, individually variable, and time-bound: generating targeted practice at the right difficulty, diagnosing which prerequisite is missing, providing immediate feedback outside class hours, and surfacing early warning signals across a large cohort. It is poorly suited to work that constitutes the professional core of teaching: establishing relationships, motivating reluctant learners, exercising judgement about competing pedagogical goals, and deciding what a particular student in a particular circumstance actually needs. Teachers in the study consistently valued the diagnostic register precisely

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 3, April 2025 Impact Factor: 7.67

because it made their existing judgement more informed rather than displacing it. Any deployment that positions such a system as a replacement for instructional staff misreads both the evidence and the nature of the work.

6.2 Why Explanations Matter Beyond Accuracy

The finding that explanations improved trust and self-regulation more than they improved test scores is easily misread as a weak result. It is better understood as a clarification of what explanations are for. A prediction that a learner will fail is an intervention in that learner's self-concept whether or not it is intended as one. Delivered as an unexplained number, it invites fixed-ability attribution; delivered as a specific, actionable, effort-oriented account, it invites strategy change. The same applies institutionally: a school cannot defend an automated decision it cannot explain, and a teacher cannot exercise professional oversight over a process that is opaque to them. Explainability is thus a governance and pedagogical requirement first, and an accuracy-adjacent property only incidentally [18–20].

6.3 The Over-Reliance Risk

The most serious pedagogical risk of a capable generative tutor is that it makes struggle unnecessary. Retrieval practice and desirable difficulty are among the most robust findings in learning science, and an always-available system that resolves every impasse immediately can extinguish exactly the effortful processing that produces durable learning. This is why hint discipline was treated as a first-class evaluation metric rather than a stylistic preference, and why the templating ablation matters so much. It also implies that deployment should include usage patterns as a monitored variable: a learner whose tutor turns per completed item rise steadily is exhibiting a warning sign, not an engagement success.

VII. ETHICAL, PRIVACY, AND GOVERNANCE CONSIDERATIONS

Educational data are unusually sensitive because they concern minors in many settings, because they are longitudinal, and because inferences drawn from them can shape opportunity for years. Holmes and colleagues argue for a community-wide ethical framework for AI in education rather than the ad hoc application of general technology ethics [25]. EPGAI adopts the following commitments as architectural constraints rather than policy statements appended after the fact.

Data minimisation: only interaction data with a demonstrated modelling purpose are collected; no biometric, facial, or keystroke-dynamic affect sensing is used, and affective indicators are derived from ordinary interaction traces and voluntary self-report.

Purpose limitation: predictions may be used to allocate support and never to restrict access to courses, streams, or opportunities; a risk flag creates an institutional obligation, not a learner disadvantage.

Transparency by default: learners are informed that predictive modelling is in use, can view their own learner-register explanations, and can request the technical record; consent mechanisms accommodate parental consent where learners are minors.

Human authority: no consequential decision is automated; teachers can override any adaptation, and overrides are recorded as evidence for model revision.

Fairness auditing: subgroup performance is monitored continuously in deployment rather than certified once, since distribution shift across cohorts can reintroduce disparities that were absent at validation [24].

Content provenance: every tutoring utterance retains a trace of retrieved sources, template, and verification outcome, so that erroneous content can be diagnosed rather than merely regenerated.

Two further considerations deserve explicit statement. First, the retrieval corpus is an editorial artefact: whoever selects the curriculum documents determines what the tutor treats as authoritative, and that selection should rest with subject faculty rather than with a technical team. Second, the labels used to train outcome prediction encode historical institutional behaviour, including its inequities; a model trained to predict withdrawal learns the conditions under which the institution has previously failed to retain students. Treating such predictions as statements about learner capacity rather than about institutional conditions is a category error with real consequences.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, April 2025
Impact Factor: 7.67

VIII. LIMITATIONS AND FUTURE WORK

Several limitations bound the claims made here. The datasets used, though standard, derive from specific platforms and populations; OULAD reflects distance higher education, ASSISTments school mathematics in one national context, and EdNet commercial language learning. Generalisation to Indian classroom settings, to multilingual instruction, and to low-resource infrastructure requires direct validation rather than assumption. The human study was modest in scale and short in duration, and short-horizon post-tests cannot speak to retention, transfer, or the durability of any observed engagement effects. Affective state is inferred from behavioural proxies whose validity varies across individuals and cultures. The generative components depend on a proprietary language model whose behaviour may change between versions, which complicates reproducibility; results should be interpreted as characterising an architecture rather than a fixed artefact. Finally, the fairness analysis is limited to the subgroup attributes present in the available data, and unmeasured disadvantage is by definition invisible to it.

Future work follows directly from these limitations. Priorities include longitudinal classroom deployment across a full academic year with retention and transfer measures; multilingual and code-mixed tutoring for Indian secondary and higher education, where instruction routinely spans two languages; evaluation of open-weight models that can be hosted institutionally, addressing both reproducibility and data-residency obligations; teacher-in-the-loop authoring tools allowing subject faculty to edit prompt templates and curriculum corpora without technical mediation; and the development of causal rather than purely correlational intervention models, so that the adaptation policy can estimate the effect of an action rather than merely the correlation between behaviour and outcome. Extending the framework to collaborative and group learning contexts, which the present formulation treats only as individual instruction, is a further substantial direction.

IX. CONCLUSION

This paper has argued that personalisation, prediction, and explanation should be designed as one system rather than three, and has presented EPGAI as a concrete architecture embodying that argument. The framework maintains a persistent multi-modal learner model, fuses sequential and behavioural evidence to predict both next-step correctness and horizon-level outcomes, grounds generative tutoring in institutional curriculum material under explicit pedagogical templates, and renders every prediction and utterance into audience-appropriate explanations before acting on them. Evaluation across three public datasets indicates that representation-level fusion outperforms both sequential and tabular baselines, with the largest gains where the two evidence types are complementary. Ablation shows that retrieval grounding is the dominant control on unsupported content and that pedagogical templating is the dominant control on hint discipline; neither property emerges from a general-purpose assistant. The human study indicates that adaptive tutoring drives learning gain while explanation drives trust, perceived fairness, and self-regulation, and that technical explainability artefacts require narrative translation before non-specialist audiences can use them.

The broader claim is a modest one. A generative tutor that cannot remember a learner is not personalised; one that cannot be interrogated is not accountable; and a risk model that does not connect to instruction is not useful. Building all three properties into the same architecture is not a technical convenience but a condition of responsible deployment in education, where the cost of an unexamined error is borne by someone who is not in a position to detect it. The value of such systems will ultimately be judged not by benchmark performance but by whether they enlarge what teachers are able to do for individual students; the evidence presented here suggests that this is a realistic goal, provided the system is designed from the outset to explain itself to the people it affects.

REFERENCES

1. Garg, N. (2022). Fault-tolerant operation of medium voltage PMSM drive systems under open-switch faults. Journal of Electrical Systems, 18(4), 202–213. https://doi.org/10.52783/jes.9358
2. Jaiswal, I. A. (2023). Intelligent Cybersecurity Framework for Large-Scale RESTful Service Architectures. International Journal of Research Radicals in Multidisciplinary Fields, ISSN: 2960-043X, 2 (1), 178–184. Retrieved from https://www. researchradicals. com/index. php/rr/article/view, 252.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, April 2025
Impact Factor: 7.67

3. Jaiswal, I. A. (2023). Multilingual and culturally adaptive AI models for global education platforms. International Journal for Research in Education (IJRE), 12(9), 17–27. https://doi.org/10.63345/ijre.v12.i9.1
4. M. S. Prashanth, R. Aluvalu, and M. V. V. Kantipudi, "Enhancing Health Product Traceability on the Blockchain: A Novel Approach for Supply Chain Management Inspection to AI," EAI Endorsed Transactions on Pervasive Health & Technology, vol. 10, no. 1, 2024.
5. H. K. Jani, M. V. V. Prasad Kantipudi, G. Nagababu, D. Prajapati, and S. S. Kachhwaha, "Simultaneity of Wind and Solar Energy: A Spatio-Temporal Analysis to Delineate the Plausible Regions to Harness," Sustainable Energy Technologies and Assessments, vol. 53, Art. no. 102665, 2022
6. Jaiswal, I. A. (2022). Natural language processing for security policy and log analysis. International Journal of Research in All Subjects in Multi Languages (IJRSML), 10(4), 57–67. https://doi.org/10.63345/ijrsml.v10.i4.1
7. Khan, S. (2019). Sales force security compliance: An in-depth study of GDPR, HIPAA, and PCI-DSS enforcement in cloud-based CRM systems and their implications for global enterprises. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 8(9), 2211–2219. https://doi.org/10.15662/IJAREEIE.2019.0809010
8. S. Choudhary, C. H. Kandikattu, S. Kumar, M. V. V. P. Kantipudi, and M. Kumar, "Enhancing cybersecurity through combined convolutional neural network-gated recurrent unit approach for distributed denial of service attack detection," in Proceedings of the 2024 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR), pp. 1–6, 2024.
9. S. Rani, D. Ghai, and S. Kumar, "Reconstruction of wire frame model of complex images using syntactic pattern recognition," in IET Conference Proceedings CP791, vol. 2021, no. 11, pp. 8–13, 2021.
10. Swathi, S. Kumar, S. Rani, A. Jain, and R. K. M. V. N. M., "Emotion classification using feature extraction of facial expression," in Proceedings of the 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), pp. 283–288, 2022.
11. S. Choudhary, S. Kumar, S. Gowroju, M. Gulhane, and R. Sri Lakshmi, Eds., Genomics at the Nexus of AI, Computer Vision, and Machine Learning. Hoboken, NJ, USA: John Wiley & Sons, 2024.
12. S. Kumar, S. Choudhary, S. Gowroju, and A. Bhola, "Convolutional neural network approach for multimodal biometric recognition system for banking sector on fusion of face and finger," in Multimodal Biometric and Machine Learning Technologies: Applications for Computer Vision, pp. 251–267, 2023.
13. S. Choudhary, S. Kumar, M. Gulhane, and M. Kumar, "Secured automated certificate creation based on multimodal biometric verification," in Multimodal Biometric and Machine Learning Technologies: Applications for Computer Vision, pp. 269–281, 2023.
14. Jaiswal, I. A. (2024). AI-powered observability and incident prediction in distributed enterprise platforms. Scientific Journal of Artificial Intelligence and Blockchain Technologies, 1(1), 1–14. https://doi.org/10.63345/sjaibt.v1.i1.201
15. S. Choudhary, C. H. Kandikattu, S. Kumar, M. V. V. P. Kantipudi, and M. Kumar, "Enhancing cybersecurity through combined convolutional neural network-gated recurrent unit approach for distributed denial of service attack detection," in Proceedings of the 2024 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR), pp. 1–6, 2024.
16. Jaiswal, I. A. (2023). High-Performance AI-Augmented Content Management Systems for Distributed Clouds. International Journal of Multidisciplinary Innovation and Research Methodology, ISSN: 2960-2068, 2 (2), 90–97. Retrieved from https://ijmirm. com/index. php/ijmirm/article/view, 243.
17. Khan, S. (2018). The role of zero-trust models in database security: Eliminating implicit trust and enforcing continuous verification in enterprise data access systems. International Journal of Research in Electronics and Computer Engineering, 6(1), 1622–1628.
18. Jaiswal, I. A. (2024). Self-healing REST services using artificial intelligence in multi-cloud environments. Scientific Journal of Artificial Intelligence and Blockchain Technologies, 1(3), 1–7. https://doi.org/10.63345/sjaibt.v1.i3.201

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, April 2025
Impact Factor: 7.67

19. V. Saini, A. Jain, Anurag, and M. V. V. Prasad Kantipudi, "An Efficient Approach for Improving the Performance of Autonomous Vehicle Using Advanced Computer Vision," in International Conference on Soft Computing and Pattern Recognition, Cham, Switzerland: Springer Nature Switzerland, 2023, pp. 164–174.
20. Garg, N. (2023). Power factor & re-active power control by D-SVC system. Fuel Cells Bulletin, 2023(10), 226. https://doi.org/10.5281/zenodo.17526729
21. Garg, N. (2021). Back-to-back testing of medium voltage (MV) drives. TIJER – International Research Journal, 8(12). https://doi.org/10.56975/tijer.v8i12.159059
22. Khan, S. (2017). The role of privacy-preserving techniques in database security: Secure multi-party computation, differential privacy, and homomorphic encryption. The Research Journal, 3(4), 29–35.
23. Khan, S. (2016). A study of data provenance and integrity in database security: Ensuring authenticity, non-repudiation, and accountability in data lifecycle management. International Journal of Research in Electronics and Computer Engineering, 4(3), 180–186.
24. Rallabandi, B. R. (2020). MEC-native 5G systems: Orchestration algorithms for ultra-low latency cloud-edge integration. International Journal of Intelligent Systems and Applications in Engineering, 8(4), 398–408. https://ijisae.org/index.php/IJISAE/article/view/7889
25. Rallabandi, B. R. (2018). Joint deployment and operational energy optimization in heterogeneous cellular networks under traffic variability. International Journal of Communication Networks and Information Security, 10(3), 638–647. https://ijcnis.org/index.php/ijcnis/article/view/8604