Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Emotion-Aware Artificial Intelligence for Personalized Learning: Integrating Affective Computing and Educational Analytics

Yumeng Duan

d1246902012cnuedu@163.com

Capital Normal University

Article

Keywords: Affective computing, Emotion-aware learning analytics, Student performance prediction, Knowledge tracing, Multimodal fusion

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11615781/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Emotion-Aware Artificial Intelligence for Personalized Learning: Integrating Affective Computing and Educational Analytics

Yumeng Duan[1*]

[1]College of Marxism, Capital Normal University, Beijing, 100080, China.

Corresponding author(s). E-mail(s): d1246902012cnuedu@163.com;

Abstract

Accurate educational prediction requires a comprehensive understanding of learners' behavioral patterns and affective states, although the limited availability of emotional observations often restricts their effective integration into predictive models. To address this challenge, we propose EA-PL, an emotion-aware educational prediction framework that incorporates an availability-aware fusion mechanism into task-specific learning representations. The framework employs separate encoders to capture learning behaviors and categorical emotional observations, using gated residual fusion to regulate the contribution of affective information to course-score prediction while supporting complementary tasks involving sequential response prediction and visual affect recognition. Experiments involving 197 student–course outcomes from 169 students, with 12,095 emotional observations, demonstrate that EA-PL achieves an MAE of 0.649 and an RMSE of 0.813, compared with 0.780 and 0.972 for the behavior-only model, while increasing $\boldsymbol{R^2}$ from 0.822 to 0.875. Further ablation and robustness experiments confirm the effectiveness of adaptive emotion fusion in improving predictive accuracy and maintaining stability under missing emotional observations. Evaluations on independent knowledge-tracing and visual affect-recognition datasets further demonstrate competitive response-prediction performance and the benefits of temporal modeling for emotion recognition.

Keywords: Affective computing; Emotion-aware learning analytics; Student performance prediction; Knowledge tracing; Multimodal fusion

1 Introduction

An incorrect response can reflect a conceptual difficulty, reduced attention or an unsuccessful attempt to resolve confusion. Response accuracy alone cannot distinguish these processes. Control–value theory relates achievement emotions to appraisals of control and task value, providing a basis for examining emotion alongside educational performance[1]. Research on affective dynamics distinguishes confusion resolved through reasoning from persistent frustration and boredom[2]. Experimental work also shows that confusion can benefit learning when it is appropriately induced and resolved[3]. These findings favor contextual interpretation over a fixed assumption that positive emotion predicts successful learning. Learning analytics examines complementary evidence from practice, task completion and study activity[4]. Deep knowledge tracing uses recurrent representations of previous responses to estimate subsequent performance[5]. Dynamic key-value memory networks represent concept-specific knowledge states[6], while self-attentive and context-aware attentive models relate a target question to relevant past interactions[7, 8]. More recent approaches include the xLSTM and item-response formulation of DKT2[9] and the cascaded cognitive stages of ACE-KT[10]. These methods predict individual responses. A final course score requires a different sample definition, prediction head and evaluation protocol.

Emotion-related datasets differ in how their labels are obtained. Video annotations describe observed learning-related affect, whereas detectors applied to interaction logs yield behavior-derived estimates. Other releases contain only categorical image-classifier outputs. DAiSEE provides ordered intensity labels for four dimensions of learning-related affect[11]. ASSISTments research has related detector-estimated affect and engagement to examination outcomes[12], and CADKT models affect-related behavioral profiles alongside cognitive structure[13]. Facial behavior provides an observable cue for affective-state recognition, and its interpretation follows established work on emotion measurement[14].

Multimodal learning distinguishes representation, alignment and fusion[15], and educational applications must account for the granularity and provenance of each source[16]. In this study, the three datasets support separate task-specific evaluations: video clips are used for affect recognition, interaction sequences for next-response prediction, and student–course records for course-score estimation. Each experiment preserves the native observation unit, participant grouping and label structure of its dataset. The public Emotion_dataset combines emotional observations, course behavior and scores[17]. Our audit identifies repeated timestamped emotions attached to a fixed behavioral vector and score within each student–course combination. The course-score task therefore combines repeated emotion observations with static course-level behavioral profiles. All courses belonging to one student are kept in the same data partition.

Validation must preserve this hierarchy. Randomly separating records can place the same behavioral vector and score in both training and test sets. Such leakage can inflate performance estimates[18]. Cross-validation should reflect the dependence structure relevant to the intended generalization[19], and model selection must remain separate from final performance assessment[20]. We therefore use the student–course combination as the prediction unit and the student as the partitioning and resampling unit. EA-PL addresses the fusion problem through a sample-dependent gate and an explicit availability mask. Multiplicative gates control the contribution of modalities in earlier fusion architectures[21]; explicit missingness representations distinguish absent observations from observed values in recurrent models[22]. The course-score model combines a behavioral encoder with categorical emotion proportions. The gate modulates the emotional update, and the availability mask suppresses that update when no valid emotion observation remains. The visual branch separately models within-clip temporal information using ordered frame features and a GRU. Availability handling and learned gating therefore perform distinct functions within the framework; probability calibration is evaluated using held-out response targets[23].

We address three connected questions in course-score prediction: whether emotional observations add predictive information beyond behavior and course context, whether gating and availability handling make distinct contributions, and how those contributions change when observations are missing or their categories are perturbed. EA-PL is evaluated on student-group partitions, with paired component comparisons, alignment controls and weighting analyses used to interpret the aggregate scores. Two independent tasks broaden the evaluation: response prediction compares discrimination and probability accuracy with recent knowledge-tracing methods, while video recognition assesses the temporal and ordinal properties of affect estimates. Together, these tasks examine educational prediction at distinct observation levels without assuming that their samples or labels are interchangeable.

2 Results

2.1 Data characteristics and observation structure

The audited workbook contains 12,095 records, 16 fields, 169 student identifiers and three course identifiers. Grouping by student and course yields 197 combinations. Twenty-eight students appear in two courses; 25 have different scores across those courses. Selecting one record per student would discard outcomes, whereas pooling emotions by student would mix observations associated with different targets. Table 1 summarizes the three evaluation settings.

Table 1 Dataset roles and observation units. Course counts are audited; DAiSEE totals follow the source documentation.

Dataset	Observation structure	Analysis unit and target	Dataset summary
DAiSEE (D1)	Video; four ordered affect dimensions	Video clip; four intensity labels	9,068 clips and 112 individuals reported in dataset documentation
ASSISTments 2017 (D2)	Responses, skills, timestamps; documented affect scores	Student interaction; next-response correctness	Sequential interaction records with documented affect scores
Emotion_dataset (D3)	Emotion records with fixed course summaries	197 student–course scores; 169 student groups	12,095 records, three courses and 16 fields

All 11 behavioral or acquisition-summary fields and the score are constant within each of the 197 student–course combinations. Only the emotion categories and timestamps vary. The dataset therefore links repeated emotional observations to static course profiles. It contains neither question identifiers nor response-correctness trajectories suitable for knowledge tracing. Scores range from 0–10, with a median of 9.0 across the 197 course samples. Observation support varies substantially (Figure 1). Each student–course combination contains one to 327 emotion records, with a median of 41. Twenty-seven combinations have fewer than ten records, including four with only one. Neutral accounts for 7,181 records, sad for 4,333 and happy for 581. These frequencies reflect the released observations; they do not estimate equally weighted student prevalence or the duration of each state.

Timestamps span 18 August to 18 September 2025. Within-course adjacent intervals have a median of 49.46 seconds, and 2,460 adjacent pairs are less than one second apart. The observations are not evenly spaced (Figure 1d).

Released emotion counts differ from reported image totals in 193 combinations; the analyses use the released emotion records and their valid-observation indicators.

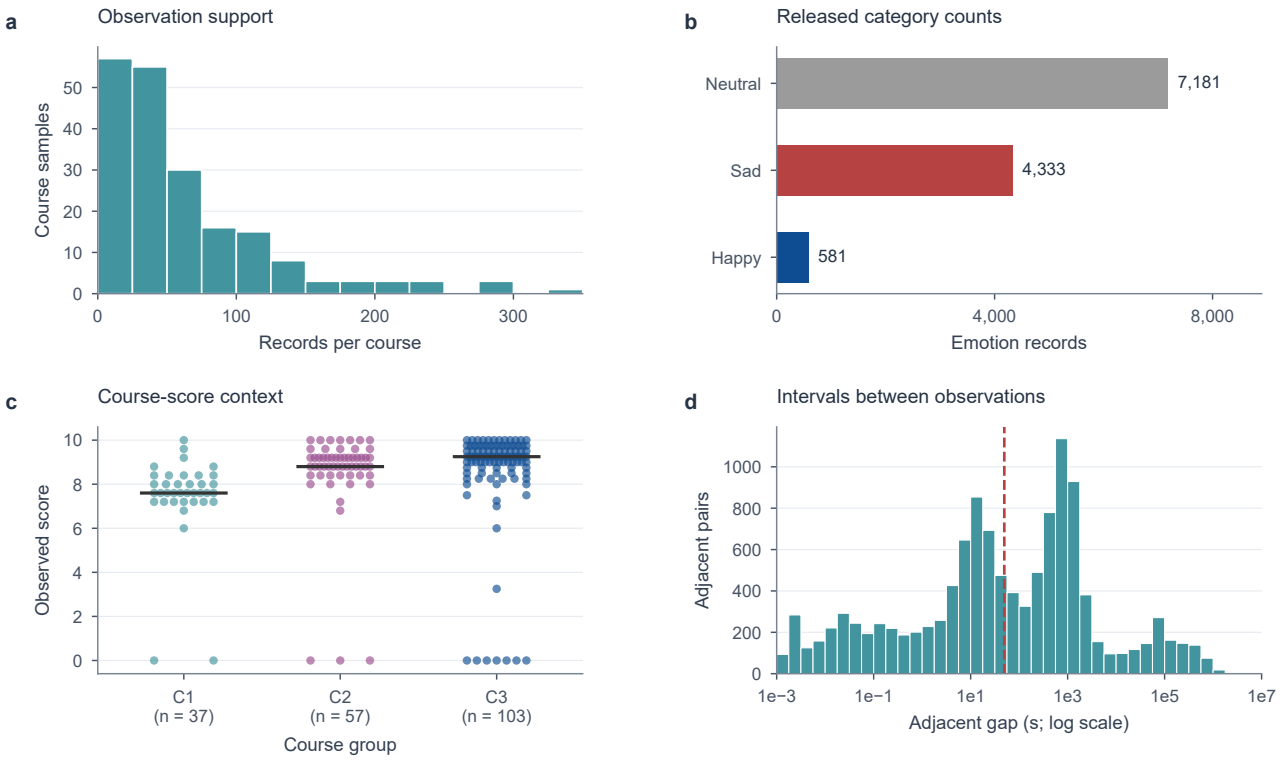

Fig. 1 Course-data structure. **a**, Record counts for 197 student–course outcomes from 169 students. **b**, Emotion frequencies across 12,095 records. **c**, Course scores ($n = 37, 57, 103$); marks denote medians and jitter separates ties. **d**, All 11,898 within-outcome time gaps on a logarithmic axis; the dashed line marks the median (49.46 s). Descriptive summaries only.

2.2 Course-score performance and component contributions

The primary evaluation makes one prediction per student–course outcome and keeps each student's courses together during partitioning. Full EA-PL reduces MAE from 0.780 for behavior-only input to 0.649 (Table 2). Because the configurations share behavioral features and course context, this comparison isolates the incremental contribution of the emotion pathway within the evaluated design. Architecturally, the residual connection retains the behavioral representation while allowing emotion to adjust it, rather than requiring the emotional branch to reconstruct the course profile.

Table 2 Clean-input course-score comparisons on 197 outcomes from 169 students. Values are mean ± sample SD over five matched random-seed runs; errors use the original score scale. Full and the unmasked variant share clean-input predictions.

Configuration	MAE ↓	RMSE ↓	R^2 ↑
Full	0.649 ± 0.028	0.813 ± 0.031	0.875 ± 0.010
Without emotion	0.780 ± 0.014	0.972 ± 0.022	0.822 ± 0.008
Without gate	0.667 ± 0.028	0.839 ± 0.030	0.867 ± 0.009
Without missingness handling	0.649 ± 0.028	0.813 ± 0.031	0.875 ± 0.010

Full has lower error than concatenation in the course-error distributions (Figure 2b). In the model formulation, the vector gate conditions the emotional update on both representations and scales its latent dimensions separately. The matched comparison supports this conditional fusion design at the aggregate prediction level. The student-level contrasts in Figure 3f describe how the error differences vary across learners. The residual distributions complement the pooled metrics by retaining the low-score tail and all course outcomes (Figure 2). Agreement across error metrics

is useful here because R^2 depends on target spread, while MAE and RMSE respond differently to large errors. The use of unclipped predictions also makes the residual analysis informative about the model itself, rather than the effect of enforcing score bounds.

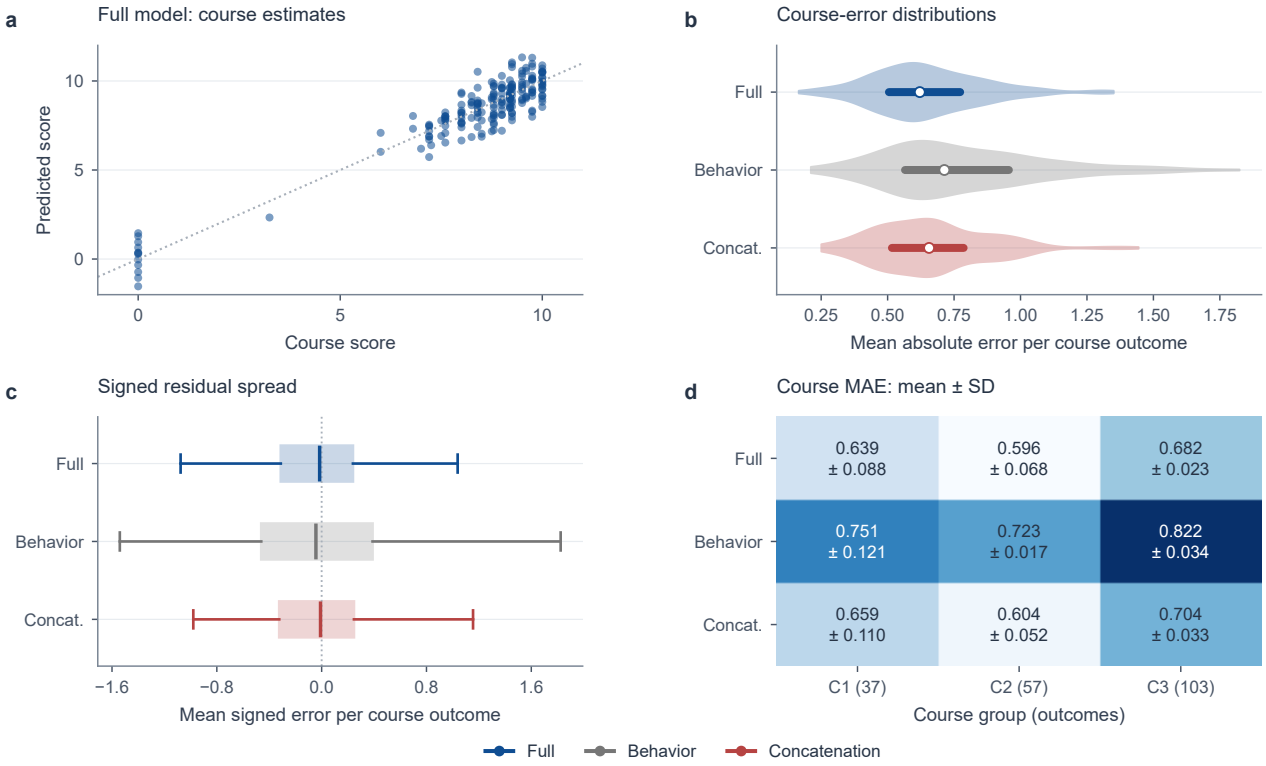

Fig. 2 Course-score predictions and errors ($n = 197$). **a**, Full predictions in the first random-seed run; the diagonal marks equality. **b**–**c**, Absolute and signed prediction errors averaged across five random-seed runs. Violins show medians and interquartile ranges; boxes show medians, interquartile ranges and full-range whiskers. **d**, Course-specific MAE (mean ± sample SD; $n = 37, 57, 103$). Predictions are unclipped; darker cells indicate higher error.

To separate the contribution of emotional input from the way it is fused, we compare Full with four configurations: without emotional input, without the gate, without missingness handling, and without both the gate and missingness handling (Figure 3a). The gate and mask are inapplicable to behavior-only input. All contrasts use the same 197 outcomes, 169 student groups and five matched random-seed runs. Table 3 reports ablated MAE minus Full MAE within each condition; positive values favor Full. The clean-input ablations separate information content from fusion control. Removing emotion produces the largest deterioration, whereas removing the gate produces a smaller penalty (Table 3; Figure 3b). Emotional input therefore contributes beyond the shared behavioral profile, and conditional scaling improves how that input is used. In contrast, mask removal leaves complete-input predictions unchanged. This is an informative control: availability handling acts when inputs are missing, rather than providing an additional clean-input correction.

Table 3 Paired MAE contrasts under matched inputs (ablated minus Full). Values are mean ± sample SD across five random-seed runs; each column uses its condition-specific Full reference. Positive values favor Full.

Removed component	Clean input	50% missing	50% category errors
Without emotion	0.131 ± 0.018	0.088 ± 0.021	0.041 ± 0.016
Without gate	0.018 ± 0.006	0.036 ± 0.003	0.026 ± 0.003
Without missingness handling	0.000 ± 0.000	0.044 ± 0.012	0.000 ± 0.000
Without gate and mask	0.018 ± 0.006	0.077 ± 0.015	0.026 ± 0.003

Under observation loss, gate and mask removal both increase error, and their joint removal gives a penalty close to the sum of the individual penalties (Figure 3c,e). The two components address different parts of the computation. The

gate controls the magnitude of an available emotional update, while missingness handling prevents unavailable observations from being processed as ordinary categories and suppresses an all-missing update. Their approximately additive penalties support complementary functions under the tested deletion condition, without requiring an additional synergistic interaction.

Category errors separate these functions more clearly (Figure 3d). Gating retains an advantage over the no-gate configuration, whereas mask removal has no effect. Because a corrupted category remains marked as available, the availability mask stays active. The result shows that the gate provides conditional scaling under perturbed emotional input, while the mask specifically handles absent observations. The student-level contrasts show that the aggregate advantage coexists with heterogeneous individual responses (Figure 3f). The mean contrast summarizes the overall benefit, while the distribution identifies variation across learners. Averaging courses within each student preserves that distinction and prevents repeated enrollment from being interpreted as independent learner evidence.

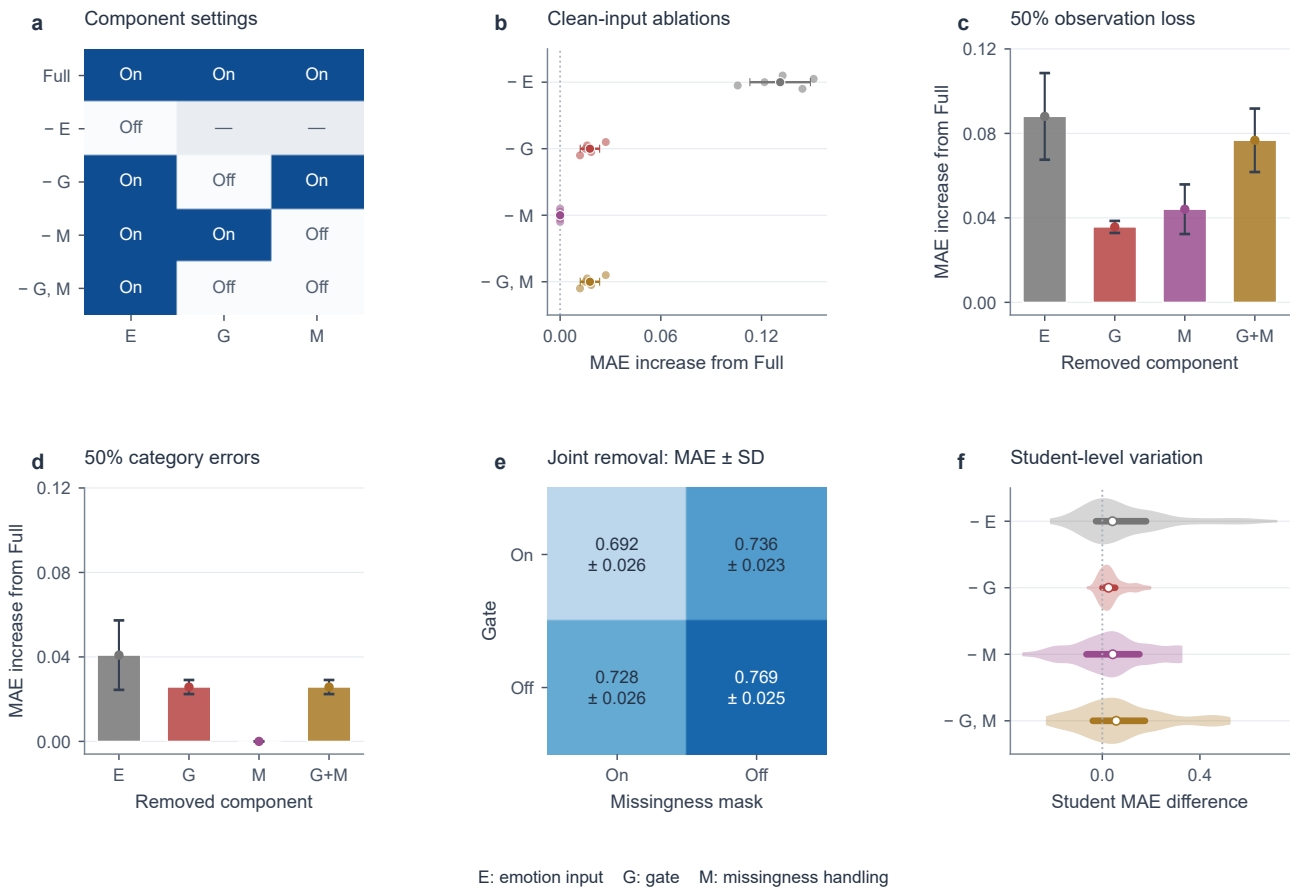

Fig. 3 Component-ablation diagnostics. **a**, Emotion (E), gate (G) and mask (M) settings; dashes indicate inapplicable components. **b–d**, Paired MAE contrasts under clean input, 50% observation loss and 50% category errors. Panel b shows all five random-seed runs; b–e summarize mean ± sample SD. **e**, Absolute MAE under 50% loss; darker cells indicate higher error. **f**, Student-level contrasts ($n = 169$), averaged over courses and random-seed runs; violins mark medians and interquartile ranges. Positive contrasts favor Full.

2.3 Robustness, alignment and sensitivity analyses

The severity analysis extends the matched ablations by tracking absolute error as input quality changes (Figure 4a). Full retains lower error than concatenation and the unmasked variant under the reported observation-loss condition. Its design explicitly separates a missing input from a valid emotion category, while retaining the behavioral residual. The comparison supports explicit availability handling under observation loss. Partial deletion also changes the category proportions estimated from the remaining observations. Full also has lower error than concatenation under the reported category-corruption condition (Figure 4b). Here, the distinction concerns fusion rather than availability: Full and the unmasked variant coincide because the observed-status indicators are unchanged. The combined perturbation analysis distinguishes the functions of the gate and availability handling. Conditional fusion remains useful with perturbed content, while explicit availability handling specifically addresses absent information.

Observation-support strata probe how the emotional representation depends on its measurement basis (Figure 4c). Proportions estimated from a short history can change substantially when an observation is removed or relabeled; longer

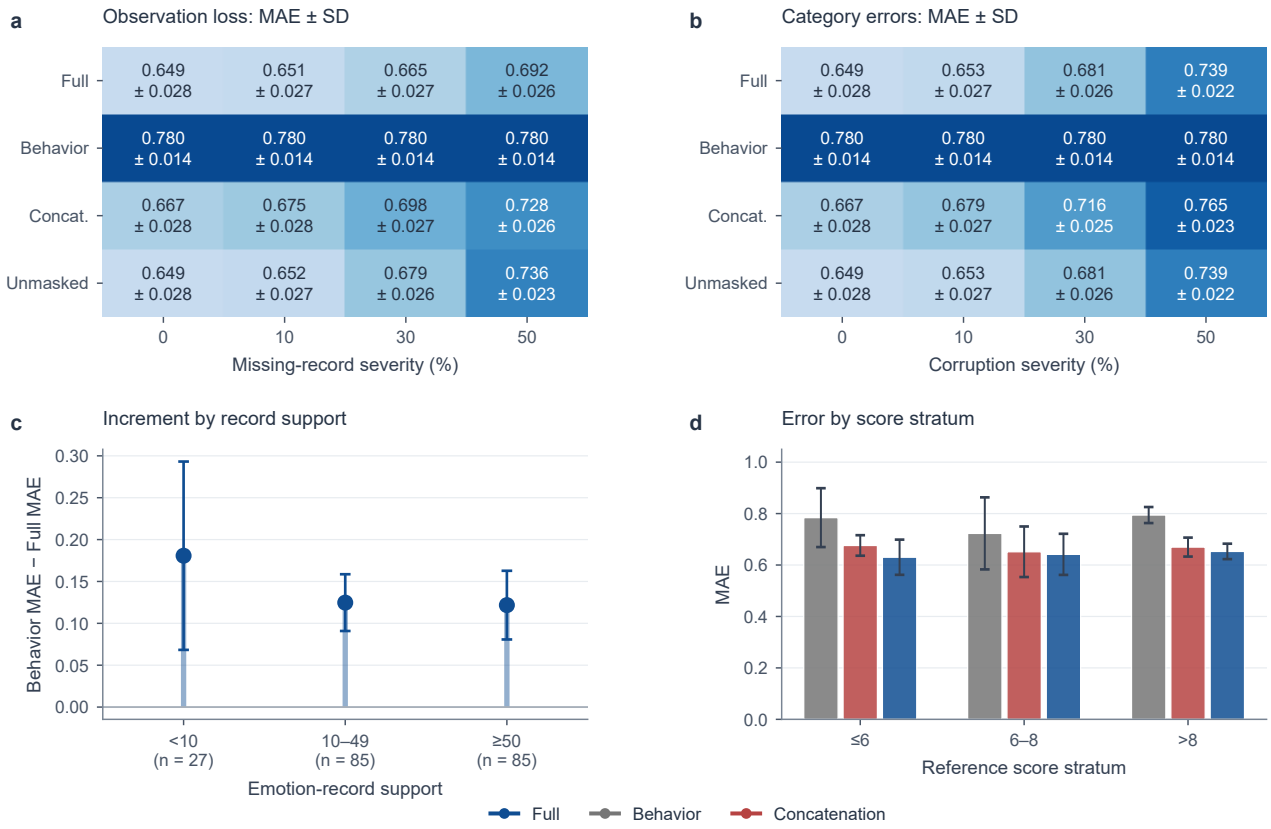

Fig. 4 Robustness and sample-stratified errors. **a**–**b**, MAE under observation loss and category corruption; darker cells indicate higher error. Full and the unmasked variant coincide in b. **c**, Behavior-only minus Full MAE by observation support; labels give group sizes. **d**, MAE by reference-score stratum: $[0, 6]$, $(6, 8]$ and $(8, 10]$. Values and error bars summarize five matched random-seed runs (mean ± sample SD).

histories provide more repeated observations but may also differ in participation and course composition. Reporting the fusion advantage within support strata helps separate the pooled comparison from these acquisition differences, without treating record count as a measure of label quality. Reference-score strata provide a complementary check on target composition (Figure 4d). Because course scores are concentrated toward the upper end of the scale, a pooled error alone can obscure performance on less frequent low-score outcomes. Retaining all outcomes and defining strata after prediction makes this analysis sensitive to that imbalance without introducing reference scores into the predictor. The pooled comparison can also be affected by course composition. Course-specific errors in Figure 2d are therefore reported alongside the aggregate metrics. These summaries locate predictive variation across courses with different observation support and sample composition.

Reassigned histories yield higher errors than the Full and behavior-only references in the within-course permutation analysis (Figure 5a). The comparison shows that the predictive value of the emotional input depends on its association with the corresponding student–course outcome. Misalignment changes the information supplied to the residual update; the resulting error distribution characterizes sensitivity to that association. The alignment and observation-support analyses characterize two complementary properties of the emotional input. History reassignment evaluates the importance of matching each emotion sequence to its student–course outcome, while support-stratified evaluation examines how prediction error varies with the number of available observations. Repeated within-course reassignments assess alignment dependence across multiple shuffles (Figure 5a). The distribution over 250 draws summarizes how errors change when the history–outcome association is disrupted while course compatibility is retained. Its spread summarizes variability across the within-course reassignments.

The correction-strength diagnostic examines whether more emotional influence is necessarily beneficial (Figure 5b). Reversing or doubling the prediction correction increases error relative to the Full endpoint, so the benefit depends on an appropriately directed and scaled adjustment. This pattern supports controlling the scale and direction of the prediction correction. The diagnostic uses behavior-only and Full prediction endpoints; it is distinct from setting the availability mask to zero inside the jointly trained model. The fusion advantage persists when each student receives equal weight after averaging their course errors (Figure 5c; Table 4). The consistent ordering across both weighting schemes supports the fusion advantage at the course and learner levels. Individual gains vary across students.

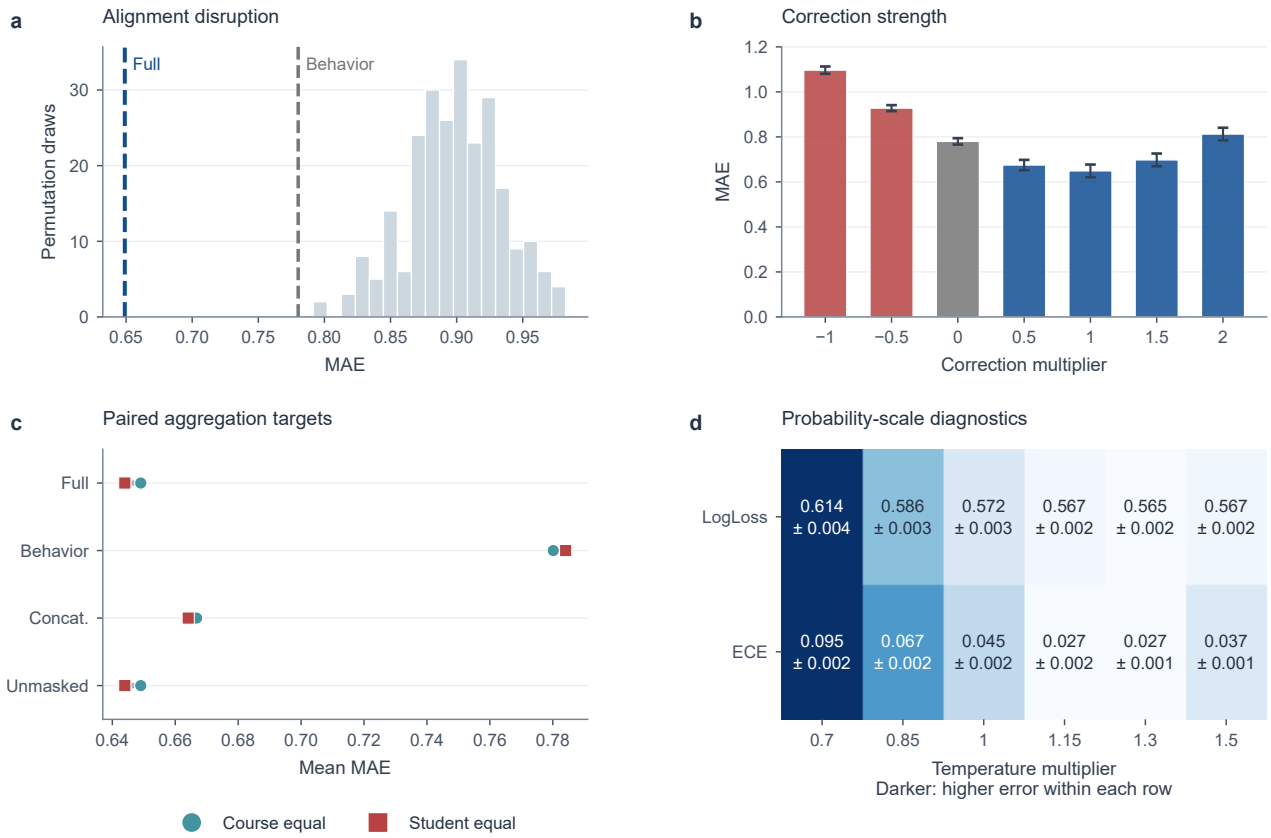

Fig. 5 Alignment and sensitivity analyses. **a**, Within-course permutation MAE (250 draws); lines mark Full and behavior-only means. **b**, Prediction-correction multipliers; 0 and 1 denote behavior-only and Full endpoints. **c**, Course-equal and student-equal MAE. **d**, Fixed-temperature response LogLoss and ECE; darker cells indicate higher error within each row. Panels b–d summarize five matched random-seed runs; bars and cells report mean ± sample SD (c: SDs in Table 4).

Course-equal MAE assigns equal weight to the 197 course outcomes. Student-equal MAE first averages course errors within each of the 169 students. The 28 students enrolled in two courses therefore receive different total weights under the two summaries, without becoming additional independent students.

Table 4 Course-equal and student-equal MAE (mean ± sample SD over five matched random-seed runs).

Configuration	Course-equal MAE	Student-equal MAE
Full	0.649 ± 0.028	0.644 ± 0.027
Without emotion	0.780 ± 0.014	0.784 ± 0.011
Without gate	0.667 ± 0.028	0.664 ± 0.027
Without missingness handling	0.649 ± 0.028	0.644 ± 0.027

2.4 Response discrimination and probability evaluation

The response task complements course-score estimation by evaluating predictions for individual interactions. Because its target is binary correctness rather than a final course score, this comparison uses a separate sequential model and response-level metrics. The response comparison uses common targets and matched random-seed runs, allowing ranking and probability error to be interpreted together (Table 5). EA-PL achieves competitive performance across both ranking and probability metrics. ACE-KT obtains the highest AUROC, while CADKT obtains the lowest LogLoss, showing complementary strengths among the evaluated knowledge-tracing methods.

The ROC and reliability curves explain why the metrics yield different orderings (Figure 6b,c). AUROC depends on the ordering of predictions, whereas calibration and probability losses also depend on their scale. EA-PL's intermediate ECE therefore identifies scope for improving probability agreement even when ranking is competitive. This distinction is relevant when predictions are used to prioritize support, because similar rankings can imply different levels of estimated risk.

Table 5 Response prediction on 9,600 targets from 240 student histories (mean ± sample SD over five matched random-seed runs). All metrics use common target–probability pairs. ECE denotes expected calibration error; arrows indicate preferred directions.

Method	AUROC ↑	Accuracy ↑	LogLoss ↓	Brier ↓	ECE ↓
DKT2[9]	0.741 ± 0.005	0.702 ± 0.004	0.573 ± 0.005	0.195 ± 0.002	0.030 ± 0.003
ACE-KT[10]	0.752 ± 0.003	0.706 ± 0.002	0.577 ± 0.003	0.195 ± 0.001	0.058 ± 0.002
CADKT[13]	0.746 ± 0.002	0.702 ± 0.001	0.570 ± 0.001	0.194 ± 0.001	0.035 ± 0.002
EA-PL	0.749 ± 0.003	0.705 ± 0.002	0.572 ± 0.003	0.194 ± 0.001	0.045 ± 0.002

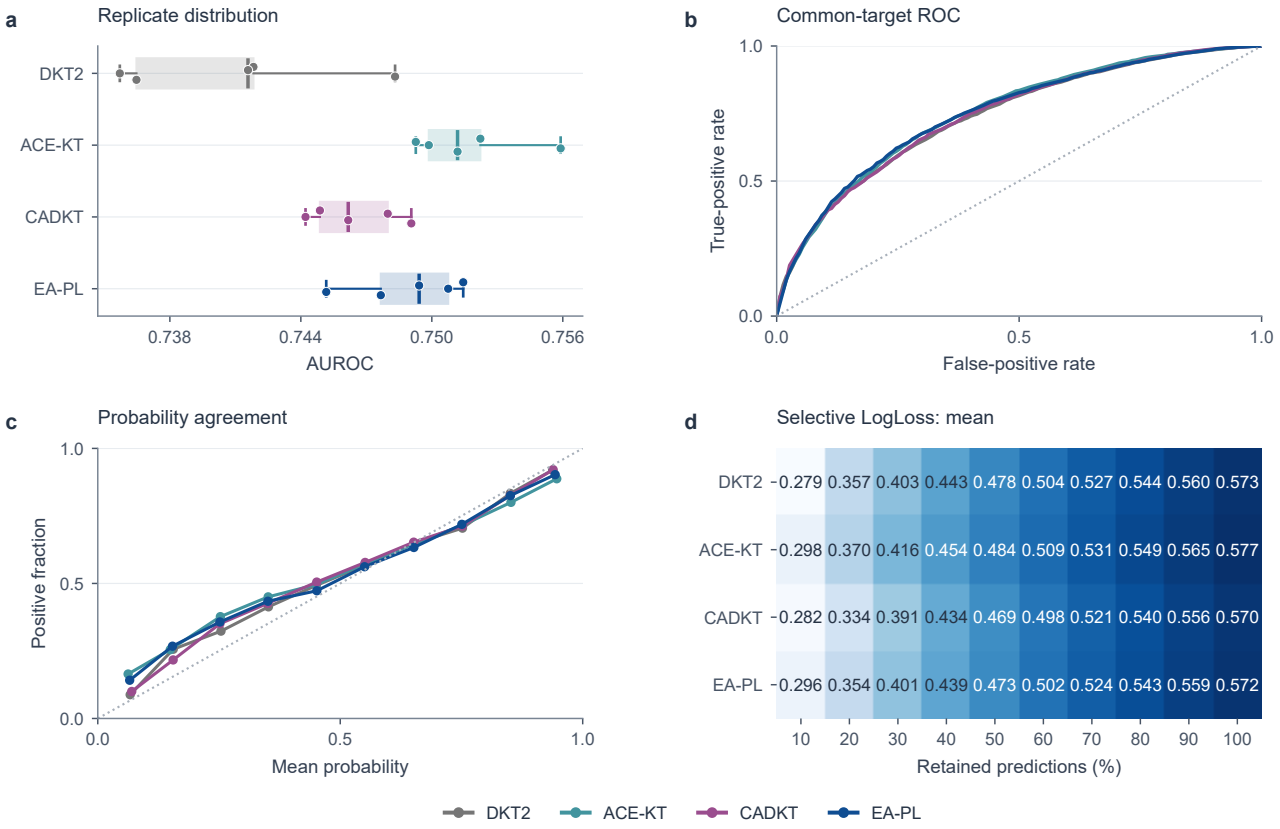

Fig. 6 Response discrimination and probability quality on 9,600 targets. **a**, AUROC across five random-seed runs; boxes show medians, interquartile ranges and full-range whiskers. **b**–**c**, First-seed-run ROC and reliability curves (ten equal-width bins; empty bins omitted). Diagonals indicate chance ranking and ideal calibration. **d**, Mean selective LogLoss across five random-seed runs, retaining 10–100% of targets by $|p - 0.5|$; darker cells indicate higher loss.

Selective prediction examines whether the probability outputs can support prioritization rather than requiring a decision on every target[24]. For EA-PL, retaining targets farther from the decision boundary lowers LogLoss relative to full coverage (Figure 6d). The result indicates that this confidence ordering identifies a lower-loss subset in the evaluated data. The analysis quantifies the trade-off between risk and coverage, with outcomes used to calculate loss after the retained subset is fixed. Temperature sensitivity separates probability sharpness from ranking quality (Table 6; Figure 5d). Moderately softening EA-PL's logits reduces probability error without changing their ordering. This response to softening identifies probability scale as a source of error independent of prediction ranking. The fixed grid quantifies sensitivity to temperature; the multipliers are specified without fitting to validation labels.

Table 6 Fixed-temperature response diagnostics (mean ± sample SD over five matched random-seed runs); multipliers are not fitted to validation labels.

Temperature multiplier	LogLoss	ECE
0.7	0.614 ± 0.004	0.095 ± 0.002
0.85	0.586 ± 0.003	0.067 ± 0.002
1.0	0.572 ± 0.003	0.045 ± 0.002
1.15	0.567 ± 0.002	0.027 ± 0.002
1.3	0.565 ± 0.002	0.027 ± 0.001
1.5	0.567 ± 0.002	0.037 ± 0.001

2.5 Comparative performance in affect recognition

The visual task addresses a measurement question distinct from either educational prediction target: how accurately can a model recognize ordered levels of learning-related affect? The evaluation includes 1,200 labeled clips per affect dimension from 30 subjects. Figure 7 compares frame-based and temporal models. Macro-F1 and unweighted average recall (UAR) assess class-balanced performance, while quadratic weighted kappa (QWK) accounts for ordinal disagreement[25].

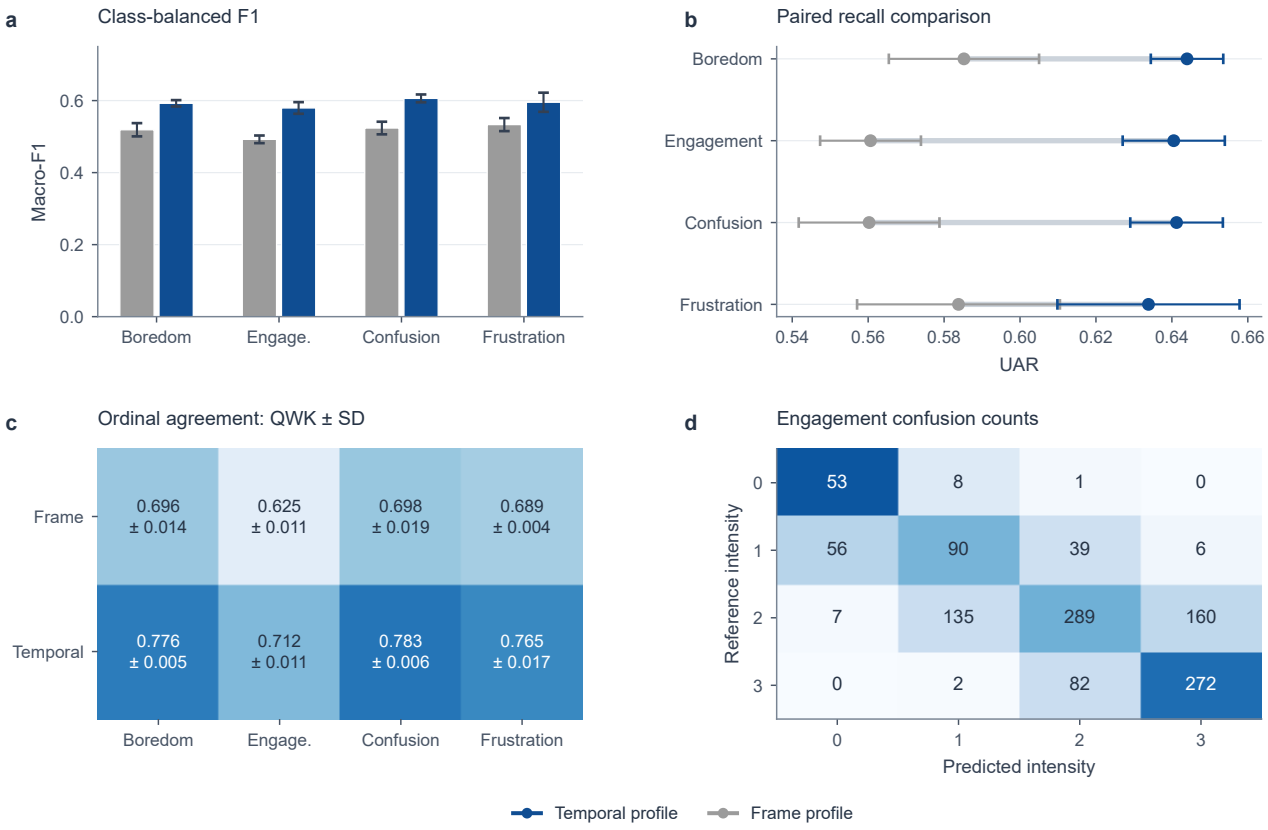

Fig. 7 Ordinal affect recognition (1,200 clips per dimension; 30 subjects). **a–c**, SigLIP 2 frame-based and GRU-based temporal Macro-F1, UAR and QWK (mean ± sample SD across five random-seed runs). Darker QWK cells indicate greater agreement. **d**, First-seed-run temporal engagement confusion counts; colors show row-normalized proportions. Labels 0–3 denote intensity levels.

The GRU-based multi-frame configuration improves class-balanced recognition over the single-frame SigLIP 2 configuration (Figure 7a,b). Both use the same spatial encoder: the former aggregates ordered representations of 16 frames, while the latter uses the midpoint frame. The comparison supports multi-frame recognition with temporal aggregation. Frame support and sequence aggregation change together in this experiment. The engagement confusion matrix concentrates predictions on the diagonal and adjacent intensity levels (Figure 7d). QWK gives nearby errors less penalty than distant errors, while Macro-F1 and UAR weight the classes equally. These metrics describe ordinal agreement and class-balanced recognition within the independent visual task.

3 Discussion

The primary finding is that emotional observations add predictive information to behavioral and course-context features in this dataset. EA-PL reduces MAE from 0.780 for behavior-only input to 0.649, with a paired reduction of 0.131 score units. The student-equal analysis preserves this ordering, showing that the improvement is consistent when each learner receives equal weight. This result is consistent with earlier work relating affect and engagement to educational outcomes[12]. The ablations distinguish three contributions. Emotional input accounts for the largest clean-input difference, gating provides a smaller reduction, and availability handling becomes relevant under observation loss. This condition dependence is central to interpreting the architecture. Gated fusion can modulate the contribution of an observed representation[21], while a mask explicitly distinguishes unavailable input from a recorded category[22]. The zero effect of mask removal on complete inputs follows its availability-dependent operation. Category corruption leaves availability unchanged, separating content perturbation from missing-input handling.

The alignment and correction-strength analyses further characterize the emotion pathway. Reassigned histories give higher errors than correctly associated inputs, and reversing or doubling the prediction correction also increases error. These findings show that the predictive contribution depends on matching each emotional history to its student–course outcome and applying the emotional update at an appropriate scale. Emotional categories, observation support and course participation jointly describe the observed learning context, while the student-level analysis shows that the size of the predictive gain varies across learners.

The complementary tasks address prediction and measurement properties separately. In response prediction, the EA-PL response branch uses interaction history and question–skill information; ACE-KT leads AUROC and CADKT leads LogLoss in the matched comparison. Temperature diagnostics characterize probability-scale sensitivity for this response branch. In visual recognition, the multi-frame GRU configuration improves engagement Macro-F1 over the single-frame reference. This contrast combines greater frame support with sequence aggregation. The three datasets retain separate learners, observation units and labels, and facial behavior supplies an observable cue whose interpretation depends on measurement context[14].

The educational contribution is improved offline course-score estimation through the integration of emotional observations and behavioral records. Course summaries and response histories operate at different levels of learning activity. The study measures predictive performance; instructional effects of recommendations constitute a separate outcome evaluated through learner interventions in intelligent tutoring research[29]. In the evaluated course data, EA-PL has lower score-prediction error than behavior-only input and ungated concatenation. The ablations distinguish the contributions of emotional input, conditional gating and missing-input handling, and perturbation analyses identify their condition-dependent behavior. Independent response and visual experiments provide complementary assessments of probability quality and ordinal affect recognition. The findings support availability-aware fusion for offline educational prediction within the reported data and evaluation setting.

Limitations. The primary dataset contains 197 outcomes from 169 students across three courses, with scores concentrated toward the upper end of the scale. Student-group validation evaluates held-out learners within this course setting[19]. Generalization across institutions, new courses and acquisition systems is a separate evaluation question. The behavioral inputs are static course summaries, so the reported task is offline score estimation. Early-warning performance depends on features available at a specified forecast time. The five-seed standard deviations describe training variability conditional on the fixed partitions. Emotional proportions summarize recorded categories rather than time spent in each state. Uneven sampling, participation and observation support can influence these summaries. The alignment and support-stratified analyses characterize these associations; a matched acquisition-only comparison would further distinguish emotional content from acquisition information. The course comparisons assess neural fusion configurations, and broader model-class comparisons require conventional tabular predictors under the same student-group evaluation. Student-level error distributions also show variation in the predictive gain across learners. The perturbation experiments use independent record deletion and uniform replacement by a different category. These operators define controlled missingness and content-error conditions. Systematic sensor failure and errors correlated with particular learners or emotional states have different structures. The availability mask responds to whether a valid observation remains; observed labels retain their influence even when their categories are incorrect. Probability diagnostics complement this input-level analysis, whereas observation noise, model uncertainty and dataset shift represent distinct evaluation dimensions[26–28].

4 Methods

Figure 8 defines the EA-PL course-score computation graph. Separate encoders represent behavior and emotion. A sample-dependent vector gate scales the emotional update, and a residual connection retains the internal behavioral pathway when emotion is unavailable. The response and video tasks use separate models.

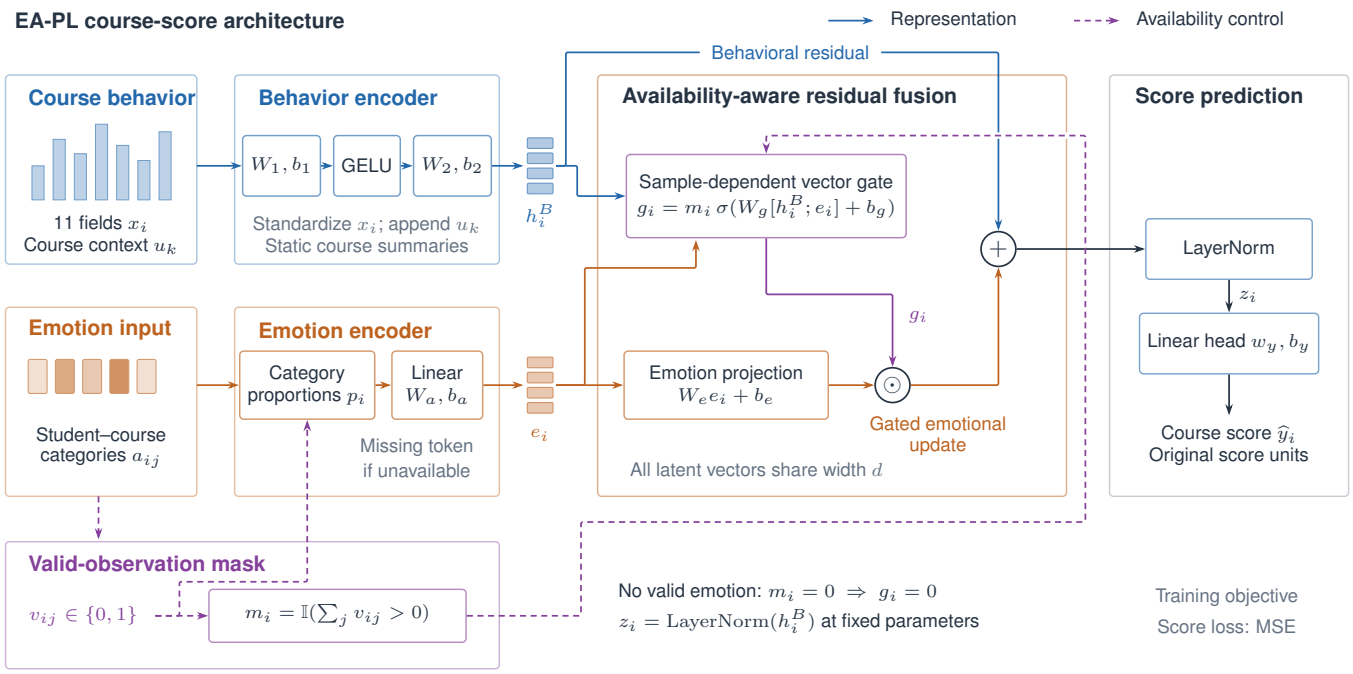

Fig. 8 EA-PL course-score model. Behavioral features and course context (blue) are fused with emotion proportions (orange) through a gated residual update, followed by layer normalization and a linear head. Validity indicators v_{ij} control aggregation; the availability mask m_i suppresses the update when no valid observation remains (purple dashed arrows). The internal behavioral pathway is retained at fixed parameters. The mask encodes availability, not label correctness; glyphs are schematic.

4.1 Study design and data preprocessing

The study has three independent evaluation arms. The primary arm estimates final course scores from static behavior and categorical emotion histories (D3). The response arm predicts next-response correctness from eligible interaction histories (D2), and the visual arm recognizes four ordinal affect dimensions from clips (D1). The primary contrasts test emotional input, gated fusion and availability handling. Alignment, correction-strength, weighting and perturbation analyses examine the sensitivity of those contrasts. Task-specific preprocessing and targets preserve the released observation structures; no students are linked across datasets. The course dataset (D3) was obtained at repository commit `82d8b89710dc9ed20ce2e20820db0e35ee368906`. A standard-library parser read the worksheet and shared-string XML. The audit counted identifiers, missing cells, within-group unique values, categories and temporal intervals. All records were included in the descriptive analysis, and none of the 16 fields contained empty cells.

DAiSEE provides ordered intensity labels for four dimensions of learning-related affect[11] and is distributed under research-use conditions. ASSISTments 2017 contains interaction sequences and documented affect scores[30]. Each dataset was evaluated as a separate task, with student or subject identities retained for grouping and evaluation rather than used as predictive attributes. The response summaries cover 9,600 next-response targets associated with 240 student histories (D2), whereas the visual summaries cover 1,200 clips from 30 subjects (D1). These are evaluation cohorts, not the full source-data totals. For evaluation at these scales, the defining sample unit remains a student history in D2 and a subject-linked clip in D1; neither collection is pooled with D3.

A course sample $i = (s, k)$ is student s in course group k. Its target is the unique `finalScore` in that group. The vector $x_i \in \mathbb{R}^{11}$ contains image count, study-session count, completed-task count, average study and completion durations, completion percentage, and five aggregate time-allocation fields. These are fixed course summaries. Group uniqueness was checked before extraction so that each student–course combination supplied one course outcome. The emotion history $\mathcal{E}_i = \{(a_{ij}, \tau_{ij})\}_{j=1}^{n_i}$ is ordered within the same student–course group. Categories have no numerical order. Relative times use the first course observation, with elapsed intervals retained for temporal analysis. All released observations enter the primary representation. The primary emotion representation uses the released records without collapsing distinct timestamps into a uniformly spaced trajectory. Training-partition means and standard deviations standardize behavioral features; zero-variance features map to zero. Scores remain in original units. Course context uses a fixed three-category encoding shared by all ablations. Student identifiers serve grouping and traceability only. Targets do not enter feature construction, and zero-score outcomes are retained in the analysis.

4.2 Representation learning and gated fusion

A small multilayer perceptron with GELU activation[31] produces the behavioral representation:

$$h_i^B = W_2\,\mathrm{GELU}(W_1[\widetilde{x}_i; u_k] + b_1) + b_2, \tag{1}$$

where $\widetilde{x}_i$ is standardized behavior and u_k is course context. Inner validation selects the latent width. The static encoder uses each course summary once instead of constructing trajectories from duplicated values.

The primary emotion encoder uses empirical category proportions:

$$p_{ic} = \frac{\sum_{j=1}^{n_i} v_{ij}\mathbb{I}(a_{ij} = c)}{\sum_{j=1}^{n_i} v_{ij}}, \qquad e_i = W_a p_i + b_a, \tag{2}$$

for a nonzero denominator, where v_{ij} indicates a valid observation. With no valid category, a designated missing token is used and fusion availability is zero. Proportions are observation-weighted, not duration-weighted. The vector p_i has one entry per released category; its denominator counts valid observations only. Observation count is treated as acquisition support rather than an emotion category.

Let $m_i = \mathbb{I}(\sum_j v_{ij} > 0)$. A vector gate scales the residual update before layer normalization[32]:

$$g_i = m_i\, \sigma(W_g[h_i^B; e_i] + b_g), \tag{3}$$

$$z_i = \mathrm{LayerNorm}\big(h_i^B + g_i \odot (W_e e_i + b_e)\big)\,, \tag{4}$$

$$\widehat{y}_i = w_y^\top z_i + b_y. \tag{5}$$

All latent vectors have the selected common width. The linear head is not clipped using test information. When $m_i = 0$, $z_i = \mathrm{LayerNorm}(h_i^B)$ exactly at fixed parameters, preserving the behavioral pathway of the jointly trained model.

The gate is learned from score error without supervision for emotion-label correctness, so its values are not calibrated reliability estimates. The primary objective is

$$\mathcal{L}_{\mathrm{score}} = \frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}}(y_i - \widehat{y}_i)^2, \tag{6}$$

where $\mathcal{T}$ contains training course samples, each once regardless of emotion-record count.

The concatenation control keeps both encoders and availability handling but replaces the gate with a learned concatenation map. The no-missingness control retains the gate, treats missing tokens as ordinary inputs, includes them in aggregation denominators and disables all-missing suppression. Training masks, computational budget and token conventions are matched. No clean-input performance difference is imposed by design.

4.3 Response prediction and affect recognition

The D2 prediction unit is the next eligible student response. An interaction is retained when it has a student identifier, problem identifier, at least one skill identifier and a binary correctness label. Events are ordered within each student by their recorded event order, with timestamps and source-row order resolving ties. Duplicate representations of the same event are removed, while distinct repeated attempts are retained in chronological order. Histories with fewer than two eligible interactions are excluded. The evaluated cohort contains 9,600 next-response targets from 240 student histories. Students are assigned to disjoint training, validation and test partitions, and all interactions from one student remain in the same partition.

For a history $\mathcal{H}_{s,t} = \{(q_\ell, k_\ell, r_\ell)\}_{\ell\le t}$, previous question–skill–response embeddings are passed through a GRU[33] to obtain $h_{s,t}^K$. A conditional prediction head uses the next question and its skill identifiers, but not its correctness:

$$\widehat{r}_{s,t+1} = \sigma\big(f_K([h_{s,t}^K; \mathrm{Emb}(q_{t+1}); \mathrm{Emb}(k_{t+1})])\big)\,. \tag{7}$$

Binary cross-entropy is evaluated on eligible next-response targets only. For a question with multiple recorded skills, the skill representation is formed by pooling the embeddings of its valid skills; unmapped question and skill identifiers use distinct unknown-token embeddings defined from training students. Training-derived vocabularies, item difficulty and normalization statistics are never refitted on validation or test targets. At every forecast cutoff, input features use the interaction prefix and known target question/skill. Target correctness and later interactions are excluded.

DKT2, ACE-KT and CADKT are the three external response-prediction baselines[9, 10, 13]. All methods use the same student-disjoint data partitions and the same 9,600 eligible response targets. Each baseline is trained with its task-specific architecture, and hyperparameters are selected on the validation students. The held-out test students are used for final evaluation. This common protocol aligns the target set, historical information and evaluation metrics across methods. The D2 branch uses interaction history, the next question and its skill identifiers as inputs. Emotion variables are not included in this response-prediction branch; emotion-aware fusion is evaluated in the course-score task, and visual affect recognition is evaluated separately. The fixed response-temperature grid divides pre-sigmoid logits by the stated positive multipliers[23]. For selective prediction, targets are ranked by $|p - 0.5|$ before loss is computed on each retained fraction.

The visual analysis is clip-level and uses four ordinal labels (0–3) per clip for boredom, engagement, confusion and frustration. A clip is included when its subject identifier, clip identifier, video and all four labels are available. The experiment contains 1,200 clips from 30 subjects. Data are partitioned by subject into training, validation and test sets, so every clip from one subject belongs to a single partition. The subject-level partition assignments remain fixed across the five random-seed runs. The frame-based visual model uses the SigLIP 2 ViT-B/16 encoder with the pretrained checkpoint `google/siglip2-base-patch16-224`. Sixteen frames are sampled uniformly from each clip and resized to 224×224 pixels using the checkpoint's input normalization. The frame-based profile uses the temporal midpoint frame, $z_i^{\text{frame}} = f_\theta(I_{i,8})$. The temporal profile applies the same frozen SigLIP 2 encoder to all 16 frames and passes the ordered frame embeddings through a one-layer GRU[33] with 256 hidden units. Its final valid hidden state is the clip representation, z_i^{temp}. The two profiles therefore share the same spatial encoder and differ in whether temporal information is aggregated across the clip.

Each visual representation is passed to four independent linear classification heads, one for each affect dimension: $\widehat{\boldsymbol{p}}_{i,d} = \text{softmax}(W_d z_i + b_d)$, $d \in \{1,2,3,4\}$. The training objective is the mean cross-entropy across the four dimensions, $\mathcal{L}_{\text{visual}} = \frac{1}{4}\sum_{d=1}^{4} \text{CE}(y_{i,d}, \widehat{\boldsymbol{p}}_{i,d})$. The SigLIP 2 encoder is frozen during classifier training; the GRU and classification heads are optimized using AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 16, a maximum of 100 epochs and early-stopping patience of 10. The selected epoch maximizes validation Macro-F1 averaged across the four affect dimensions. Test subjects are evaluated once in each random-seed run. Macro-F1, unweighted average recall and quadratic weighted kappa are calculated independently for each dimension.

4.4 Model optimization and evaluation procedures

D3 uses five fixed, outer student-group folds, preserving all courses of a student within one fold; four inner student-group folds select configurations on validation MAE. Every course outcome is tested only when its student is in the held-out outer fold. All feature standardization and any learned preprocessing are fitted using training students within the relevant fold. The specified search considers latent widths 16 and 32 and learning rates 3×10^{-4} and 10^{-3}. Training uses AdamW[34], weight decay 10^{-4}, batch size 32, dropout 0.1, at most 200 epochs and early-stopping patience 20, with stopping and hyperparameters determined without outer-test outcomes.

All experiments are repeated with five different random seeds while keeping the training, validation and test assignments fixed. The same partition definitions and eligible samples are used for every seed and every compared method. Random seeds control parameter initialization, minibatch order and stochastic training operations. For D3, out-of-fold predictions from the five outer student-group folds are pooled within each seed to calculate MAE, RMSE and R^2. The reported mean and sample standard deviation are calculated across the five seed-level metrics. Paired ablations use identical student–course outcomes within each seed. D1 keeps subject assignments fixed, and D2 keeps student assignments fixed, across the five runs. The same five-seed protocol is applied to the visual, response-prediction and course-score experiments. All reported means and sample standard deviations are calculated from the corresponding five runs under fixed data partitions.

Course metrics use residuals $\epsilon_i = y_i - \widehat{y}_i$:

$$\text{MAE} = N^{-1}\sum_i |\epsilon_i|, \quad \text{RMSE} = \sqrt{N^{-1}\sum_i \epsilon_i^2}, \quad R^2 = 1 - \frac{\sum_i \epsilon_i^2}{\sum_i (y_i - \overline{y})^2}. \tag{8}$$

The paired contrast is behavior-only MAE minus Full MAE. Component effects use paired, within-seed differences on the same test targets before calculating their mean and sample SD across seeds. Course samples receive equal weight unless student weighting is explicitly stated; grouping information is retained when analyzing student-level variation.

Response evaluation uses AUROC[35], accuracy at threshold 0.5, binary LogLoss, Brier score $N^{-1}\sum_i (p_i - y_i)^2$[36], and ECE. ECE uses ten fixed equal-width bins and weights the absolute difference between mean probability and positive fraction by bin size. ROC and reliability curves use the first random-seed run; tables summarize all five seed runs. Selective-risk and sensitivity displays aggregate seed runs as defined in their captions. Macro-F1 averages the four classwise F1 values, UAR averages classwise recall, and QWK uses squared distance weights $(a-b)^2/9$ for ordinal labels 0–3.

Error bars and mean-plus-or-minus-SD annotations summarize variability across five matched random-seed runs using fixed data partitions; sample SD is computed over runs rather than over cross-validation folds. AUROC boxes show quartiles and ranges across seed runs. Course-error and student-level ablation violins show smoothed distributions over the stated units, with medians and interquartile ranges. The permutation histogram contains 250 draws and uses the stated central quantile range. Selective LogLoss ranks targets by $|p - 0.5|$ before retaining a fixed fraction. Course-error distributions use one mean absolute error per outcome; student-equal MAE first averages courses within each student, then averages the 169 students. These summaries are descriptive. Run-to-run standard deviations are not student-cluster confidence intervals.

4.5 Ablation, perturbation and sensitivity protocols

Observation-loss evaluation uses per-record deletion probabilities $q \in \{0, 0.1, 0.3, 0.5\}$, matching the conditions in Figure 4a. Within each student–course history, a Bernoulli(q) indicator is sampled independently for every valid emotion record, and the category and timestamp are removed together. The realized deletion fraction varies across histories. If a history becomes empty, the availability indicator is zero and no neutral category is imputed. The resulting suppression of the emotional update follows the fusion equations at fixed model parameters. For each severity, 20 independently sampled deletion-mask sets are generated over the held-out student–course samples. Each mask set is shared by Full and the corresponding ablation models and is reused across the five random-seed runs. MAE is calculated for each seed–mask combination, averaged over the 20 masks within each seed, and then summarized as mean ± sample SD across the five seed-level means. The original train, validation and test partitions remain fixed, and deletion masks are applied only during evaluation.

Alignment checks reassign complete emotion histories within compatible course-enrollment groups while preserving each history as a complete sequence. Observation-support analyses stratify predictions by the number of valid emotion records. Full, behavior-only, ungated concatenation, gating without availability handling, and joint gate–mask removal are evaluated on the same student–course outcomes. Within each training seed, paired absolute-error contrasts are formed before calculating the mean and standard deviation across seeds. Results are separated by clean input, observation deletion and category corruption rather than combining their non-equivalent mechanisms. For category corruption, each available emotion record is independently selected with probability $q \in \{0, 0.1, 0.3, 0.5\}$ and its category is replaced uniformly by one of the other two released categories. Its timestamp, sequence position and availability flag remain unchanged. For each q, 20 independent corruption-mask and replacement sets are shared across the compared models and five random-seed runs. MAE is averaged over the 20 corruption draws within each seed, and the five seed-level means are summarized as mean ± sample SD.

The correction-strength diagnostic scales the prediction difference between the Full and behavior-only configurations. For paired predictions, the diagnostic is $\widehat{y}_i(\alpha) = \widehat{y}_i^B + \alpha(\widehat{y}_i^F - \widehat{y}_i^B)$, where $\widehat{y}_i^B$ and $\widehat{y}_i^F$ denote behavior-only and Full predictions on the same outcome. Multipliers 0 and 1 recover these endpoints; negative multipliers reverse the correction and multipliers above 1 extrapolate it. Model parameters remain fixed. This prediction-level diagnostic is distinct from setting the availability mask to zero inside the jointly trained model. Course-specific, record-support and reference-score strata are formed after prediction. Course-equal MAE weights the 197 outcomes equally; student-equal MAE first averages courses within the 169 students. Perturbation draws remain paired evaluations of these same outcomes.

Data availability. The audited Emotion_dataset is available at https://github.com/TranUoc/Emotion_dataset under the repository's academic-research conditions. DAiSEE and ASSISTments 2017 require official registration or application at https://people.iith.ac.in/vineethnb/resources/daisee/index.html and https://sites.google.com/view/assistmentsdatamining/dataset. The datasets were used for separate task-specific evaluations.

Code availability. The code for data preprocessing, model implementation, training and evaluation is available from Yumeng Duan upon reasonable request.

Author contributions. Yumeng Duan conceived and designed the study, prepared the data, performed the analyses and wrote the manuscript.

Competing interests. The author declares no competing interests.

Funding. This research received no external funding.

References

[1] Pekrun, R. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review* **18**, 315–341 (2006).

[2] D'Mello, S. & Graesser, A. Dynamics of affective states during complex learning. *Learning and Instruction* **22**, 145–157 (2012).

[3] D'Mello, S., Lehman, B., Pekrun, R. & Graesser, A. Confusion can be beneficial for learning. *Learning and Instruction* **29**, 153–170 (2014).

[4] Romero, C. & Ventura, S. Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery* **10**, e1355 (2020).

[5] Piech, C. *et al. Deep knowledge tracing.* In *Advances in Neural Information Processing Systems*, Vol. 28, 505–513 (2015). URL https://proceedings.neurips.cc/paper/2015/hash/bac9162b47c56fc8a4d2a519803d51b3-Abstract.html.

[6] Zhang, J., Shi, X., King, I. & Yeung, D.-Y. *Dynamic key-value memory networks for knowledge tracing.* In *Proceedings of the 26th International Conference on World Wide Web*, 765–774 (2017). URL https://archives.iw3c2.org/www2017/proceedings/proceedings/p765.pdf.

[7] Pandey, S. & Karypis, G. A self-attentive model for knowledge tracing (2019). URL https://arxiv.org/abs/1907.06837. arXiv:1907.06837.

[8] Ghosh, A., Heffernan, N. & Lan, A. S. Context-aware attentive knowledge tracing (2020). URL https://arxiv.org/abs/2007.12324. arXiv:2007.12324.

[9] Zhou, Y., Han, W. & Chen, J. Revisiting applicable and comprehensive knowledge tracing in large-scale data (2025). URL https://arxiv.org/abs/2501.14256. Author implementation: https://github.com/zyy-2001/DKT2, arXiv:2501.14256.

[10] Guo, T. *et al. ACE-KT: Cascaded cognitive modeling for stage-wise knowledge tracing.* In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, Vol. 300 of *Proceedings of Machine Learning Research*, 4537–4545 (PMLR, 2026). URL https://proceedings.mlr.press/v300/guo26b.html.

[11] Gupta, A., D'Cunha, A., Awasthi, K. & Balasubramanian, V. DAiSEE: Towards user engagement recognition in the wild (2022). URL https://arxiv.org/abs/1609.01885v7. Version 7; first released in 2016, arXiv:1609.01885.

[12] Pardos, Z. A., Baker, R. S. J. D., San Pedro, M. O. C. Z., Gowda, S. M. & Gowda, S. M. Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytics* **1**, 107–128 (2014).

[13] Wang, H., Qiu, L. & He, M. CADKT: A cognitive-affective dual-graph contrastive framework for dynamic affect-aware knowledge tracing. *Electronics* **15**, 3433 (2026). URL https://www.mdpi.com/2079-9292/15/15/3433.

[14] Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M. & Pollak, S. D. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest* **20**, 1–68 (2019).

[15] Baltrušaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: A survey and taxonomy (2017). URL https://arxiv.org/abs/1705.09406. arXiv:1705.09406.

[16] Chango, W., Lara, J. A., Cerezo, R. & Romero, C. A review on data fusion in multimodal learning analytics and educational data mining. *WIREs Data Mining and Knowledge Discovery* **12**, e1458 (2022). URL https://arxiv.org/abs/2511.20871.

[17] TranUoc. Emotion_dataset: Learning behavior and emotion data for academic performance prediction (2026). URL https://github.com/TranUoc/Emotion_dataset. Repository identifier; data collected in 2025; audited commit 82d8b89710dc9ed20ce2e20820db0e35ee368906 dated 9 May 2026; accessed 27 October 2026.

[18] Kapoor, S. & Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **4**, 100804 (2023).

[19] Roberts, D. R. *et al.* Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017).

[20] Cawley, G. C. & Talbot, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* **11**, 2079–2107 (2010). URL https://www.jmlr.org/papers/v11/cawley10a.html.

[21] Arevalo, J., Solorio, T., Montes-y Gómez, M. & González, F. A. Gated multimodal units for information fusion (2017). URL https://arxiv.org/abs/1702.01992. arXiv:1702.01992.

[22] Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports* **8**, 6085 (2018).

[23] Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. *On calibration of modern neural networks.* In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 of *Proceedings of Machine Learning Research*, 1321–1330 (PMLR, 2017). URL https://proceedings.mlr.press/v70/guo17a.html.

[24] Geifman, Y. & El-Yaniv, R. *SelectiveNet: A deep neural network with an integrated reject option.* In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, 2151–2159 (PMLR, 2019). URL https://proceedings.mlr.press/v97/geifman19a.html.

[25] Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**, 213–220 (1968).

[26] Kendall, A. & Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? (2017). URL https://arxiv.org/abs/1703.04977. arXiv:1703.04977.

[27] Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles (2017). URL https://arxiv.org/abs/1612.01474. arXiv:1612.01474.

[28] Ovadia, Y. *et al.* Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift (2019). URL https://arxiv.org/abs/1906.02530. arXiv:1906.02530.

[29] Ma, W., Adesope, O. O., Nesbit, J. C. & Liu, Q. Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology* **106**, 901–918 (2014).

[30] ASSISTments Data Mining Competition. The 2017 ASSISTments datamining competition: Dataset and column descriptions (2017). URL https://sites.google.com/view/assistmentsdatamining/dataset. Official documentation; accessed 27 October 2026.

[31] Hendrycks, D. & Gimpel, K. Gaussian error linear units (GELUs) (2016). URL https://arxiv.org/abs/1606.08415. arXiv:1606.08415.

[32] Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer normalization (2016). URL https://arxiv.org/abs/1607.06450. arXiv:1607.06450.

[33] Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling (2014). URL https://arxiv.org/abs/1412.3555. arXiv:1412.3555.

[34] Loshchilov, I. & Hutter, F. *Decoupled weight decay regularization.* In *International Conference on Learning Representations* (2019). URL https://arxiv.org/abs/1711.05101.

[35] Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861–874 (2006).

[36] Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1–3 (1950).