AI-Powered Adaptive Assessment Using NLP-Based Scoring and Sentiment-Driven Item Selection for Personalized Learning

Bryan Daniel Angelino[1,*], Ariel Christopher Wawolangi[2]

[1,2]Department of Information Systems, Faculty of AI and Data Science, Universitas Pelita Harapan, Indonesia

ABSTRACT

This research proposes and evaluates an AI-powered adaptive assessment framework that integrates automatic scoring and sentiment analysis to personalize item sequencing in digital learning environments. The study utilizes a dataset of 14,960 open-ended student responses sourced from 420 learners across 48 assessment items, each labeled with a six-level cognitive score and three-level sentiment polarity. A transformer-based scoring model achieved an exact-match accuracy of 0.64, a macro F1-score of 0.71, and a quadratic weighted kappa of 0.82, while correlating strongly with expert ratings (Spearman = 0.87). The sentiment classifier reached an overall accuracy of 0.79, with a macro F1-score of 0.76, and demonstrated recall of 0.83 for neutral responses, 0.72 for negative responses, and 0.69 for positive responses. Simulation of 50 adaptive learners over 12 steps showed an average difficulty progression of +1.8 bands, a 96 percent overload-prevention rate, and sentiment-triggered difficulty downgrades in 14 percent of selections, stabilizing difficulty variation in 72 percent of learners by step eight. These results confirm that NLP-based scoring reliably models expert judgment, sentiment inference identifies affective modulation, and the joint use of both signals enables real-time, emotion-aware adaptation. The findings demonstrate that open-ended assessment can be automated with cognitive-affective sensitivity, providing scalable personalization for online learning ecosystems.

Keywords AI-based assessment, Natural Language Processing, automated scoring, sentiment analysis, adaptive learning

Introduction

The shift toward digital and personalized learning environments has increased the demand for assessment models capable of evaluating students rapidly, fairly, and with minimal human oversight. Conventional rubric-based scoring, particularly for open-ended responses, is labor-intensive and often inconsistent across raters, creating variability in instructional feedback and slowing learning cycles [1], [2]. At the same time, most online assessments continue to operate with static difficulty sequencing every student receives the same items in the same order regardless of performance resulting in inefficient measurement and limited diagnostic value [3], [4]. These structural challenges create a growing need for assessment systems that are scalable, automated, and capable of adapting based on learner behavior.

In response to these limitations, Natural Language Processing (NLP) has become a promising mechanism for evaluating free-text responses at scale. Transformer-based language models are increasingly being used to estimate semantic quality and content accuracy, demonstrating strong performance in educational scoring tasks across multiple subject domains [5], [6], [7]. However,

Submitted: 10 January 2025
Accepted: 18 February 2025
Published: 1 November 2025

*Corresponding author
Bryan Daniel Angelino,
1081230011@student.uph.edu

Additional Information and Declarations can be found on page 330

How to cite this article: B. D. Angelino, A. C. Wawolangi, "AI-Powered Adaptive Assessment Using NLP-Based Scoring and Sentiment-Driven Item Selection for Personalized Learning," *Adapt. Learn.*, vol. 1, no. 4, pp. 314-332, 2025.

most existing NLP-based evaluation frameworks focus solely on cognitive accuracy and overlook emotional cues embedded in student language. As a consequence, they cannot capture frustration, confusion, confidence, or resignation affective variables that strongly influence persistence, cognitive load, and learning gains [8], [9]. This cognitive-only approach represents a significant gap in assessment research.

In addition, machine-graded systems generally do not influence subsequent instructional decisions in real time. While adaptive testing has been explored in psychometrics through Item Response Theory and Computerized Adaptive Testing, these systems assume structured item formats such as multiple-choice or numerical response items and do not support open-ended conceptual reasoning [10], [11]. They also omit affective considerations, treating emotional fluctuation as noise rather than a pedagogical signal [12]. Therefore, current adaptive engines cannot distinguish between lack of knowledge and emotional disengagement, nor can they modulate difficulty to maintain motivation.

The present research addresses these gaps by integrating automatic scoring with sentiment analysis to produce a dual-signal adaptive mechanism. The core objective is to build a system that simultaneously predicts cognitive correctness and affective polarity, using both outputs to regulate item difficulty. This dual-signal approach enables the assessment engine to accelerate difficulty when confidence and mastery are high, but reduce cognitive pressure when emotional negativity indicates potential disengagement [13], [14]. The purpose of this work is not merely to automate scoring but to optimize the learning trajectory itself through emotion-aware adaptation.

The novelty of this study lies in combining transformer-based scoring models with sentiment classifiers as active decision parameters in an adaptive item-selection policy. Prior research in affective computing has examined emotional recognition in educational settings, yet these signals are typically used for dashboard visualization rather than direct instructional control [15], [16]. Similarly, studies in automated scoring have evaluated model accuracy but have not operationalized scoring outputs for real-time adaptation [17]. By merging these strands, this research introduces an adaptive framework that treats text sentiment as an actionable input rather than a descriptive outcome.

Furthermore, the study contributes a methodological innovation by applying multi-signal adaptation to open-ended text an area conventionally resistant to automation. Unlike multiple-choice sequencing, open text responses require semantic interpretation, stance differentiation, and contextual reasoning. Incorporating sentiment signals adds another layer of nuance, allowing the system to distinguish between low performing yet optimistic students and high performing but frustrated ones two learner profiles that require different pedagogical responses [18], [19], [20].

Through this integration, the study aims to demonstrate that automated scoring, when coupled with emotional inference, can improve personalization, reduce overload, and sustain learner engagement. The ultimate research objective is to generate an adaptive assessment system that operates with minimal human scoring effort while maximizing diagnostic sensitivity. In doing so, the study addresses existing methodological gaps, proposes a novel computational architecture, and supports scalable, data-driven personalization in online

learning ecosystems [21], [22], [23].

Literature Review

Contemporary research on AI-enabled assessment increasingly converges on two complementary agendas: (1) scaling evaluation of open-ended student work using NLP, and (2) shifting from static testing toward adaptive and personalized measurement. Within the NLP agenda, transformer-based language models have become the dominant paradigm for representing student responses because they encode contextual semantics and reduce dependence on manual feature engineering, enabling more robust scoring of short explanations, arguments, and reflective answers across domains [11], [12]. This capability is especially relevant for formative contexts where timely feedback is as important as grading accuracy, since automated scoring can shorten feedback loops and support iterative learning cycles in digital classrooms [13]. However, the literature also emphasizes that the reliability and validity of automated scoring depend on careful alignment with rubrics, transparent evaluation protocols, and attention to bias and fairness across student subgroups [14], [15].

A parallel stream of work examines how assessment quality can be improved by integrating psychometric principles with machine learning. Classical approaches in computerized adaptive testing optimize item selection using information functions and latent proficiency estimates, typically under Item Response Theory assumptions [16]. Yet, conventional adaptive testing is historically optimized for structured responses and often under-specifies how to incorporate open-ended reasoning, discourse quality, or semantic correctness. Recent studies thus explore hybrid approaches that combine representation learning with adaptive policies, aiming to retain psychometric rigor while enabling richer response formats [17], [18]. This creates a methodological gap that motivates architectures capable of scoring constructed responses and using those scores directly as decision signals for adaptive sequencing.

Affective computing and learning analytics add a third essential dimension: students' emotional and motivational states materially affect persistence, cognitive load, and performance trajectories in online learning. Empirical studies consistently show that affective signals—such as frustration, boredom, anxiety, and confidence—predict disengagement and dropout risk, particularly in self-paced or low-supervision environments [19], [20]. Consequently, sentiment analysis and broader emotion detection methods are increasingly used to infer learner states from text, clickstream, and interaction traces, supporting early warning and intervention designs [21]. Despite these advances, many systems still treat affective outputs as descriptive analytics rather than operational variables for real-time instructional control, leaving a gap between affect detection and adaptive action [22].

Within sentiment analysis itself, the education domain imposes distinctive constraints compared with generic social media or product review settings. Student responses often mix factual exposition with subtle evaluative language, and polarity may be expressed indirectly through hedging, uncertainty markers, or rhetorical questions. This raises challenges for label consistency and model generalization, motivating domain adaptation, carefully constructed annotation guidelines, and evaluation beyond overall accuracy to include class-level precision/recall and confusion patterns [23]. Furthermore, the literature warns

that sentiment and emotion models can be sensitive to cultural and linguistic variation, which is particularly relevant in multilingual learning contexts and in educational environments where formal language norms differ from everyday discourse [24].

Recent methodological trends therefore point toward multi-task learning and joint modeling strategies, where a shared encoder supports both cognitive scoring and affect inference. Multi-task approaches can improve sample efficiency and representation quality by exploiting shared linguistic structure, while separate task heads allow the model to learn distinct decision boundaries for correctness and affect [25]. However, prior work often evaluates these tasks independently and stops short of integrating both outputs into an explicit adaptive policy. This creates an open design space for "dual-signal" adaptive assessment, where proficiency and affect jointly regulate item difficulty, pacing, and supportive feedback strategies, with the objective of maximizing measurement efficiency while minimizing overload and disengagement [26].

Methodology

Research Design

This study adopts a quantitative experimental design to develop and evaluate an AI-powered adaptive assessment system that leverages NLP and sentiment analysis. The system is designed to automatically score open-ended responses, infer affective states from student language, and adapt subsequent items based on both cognitive performance and sentiment signals. The overall workflow consists of four main stages: data collection, text preprocessing, model training, and adaptive assessment deployment.

The research is conducted in two main phases. The first phase focuses on building and validating the NLP and sentiment analysis models using historical response data and expert-annotated labels. The second phase evaluates the adaptive assessment mechanism in a simulated or real classroom setting, where students interact with the system and their learning trajectories are monitored. In both phases, performance metrics such as accuracy, F1-score, RMSE for scoring, and user experience indicators are collected.

To support transparency and reproducibility, the research design is documented in a methodological flow diagram and a summary table of key stages and artifacts. The flow from raw text input to adaptive decision-making is visualized in a process diagram, while dataset properties, annotation schemes, and evaluation metrics are tabulated.

Figure 1 visualizes the end-to-end workflow of the AI-powered adaptive assessment system. The diagram starts with the collection of student responses, which are then passed through a text preprocessing and NLP encoding stage. This block encapsulates all linguistic transformations applied to raw text, including tokenization, normalization, and transformer-based embedding generation. The output of this stage is a set of dense semantic representations that are suitable for downstream machine learning models.

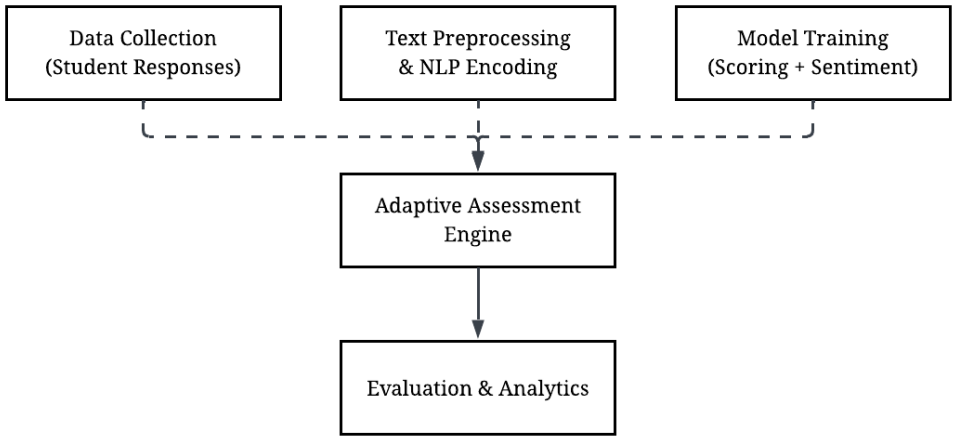

Figure 1 Research methodology flow for AI-powered adaptive assessment.

The next block represents model training, where a shared encoder feeds into two task-specific heads responsible for scoring and sentiment analysis. Predictions from these models are then consumed by the adaptive assessment engine, which dynamically adjusts item difficulty and selection strategies based on both cognitive performance and affective signals. Finally, the evaluation and analytics block close the loop by monitoring system performance, learner trajectories, and overall effectiveness. This flow clarifies how data, models, and decision policies interact in a coherent pipeline.

Dataset and Data Collection

The dataset consists of students' open-ended responses to assessment items in a specific subject domain (e.g., reading comprehension, argument writing, or conceptual explanations in STEM). Each response is paired with an expert-generated score that reflects cognitive performance (e.g., 0–5 rubric) and a sentiment label that reflects affective tone (e.g., negative, neutral, positive, or fine-grained polarity). Additional metadata such as student ID (anonymized), item ID, timestamp, and assessment session are also stored to support longitudinal analysis and adaptive sequencing.

Data collection is conducted in collaboration with educational institutions or online learning platforms. Students' complete assessments in a controlled environment where they are informed about the research and data usage. Responses are captured through a web-based interface and stored in a secure database. A subset of responses is manually annotated by domain experts and trained raters for both content accuracy and sentiment, providing a gold standard for model training and evaluation.

Descriptive statistics of the dataset are computed to understand its distribution and potential biases, including the balance of scores across proficiency levels and the distribution of sentiment labels. This information is used to guide sampling strategies, data augmentation, and model calibration to avoid overfitting on dominant classes and to ensure fair performance across the range of student abilities and emotional expressions.

Table 1 summarizes the main descriptive statistics of the dataset used to train

and evaluate the models. The table shows the number of students, items, and total responses, which together describe the overall scale of the study. The score scale and score distribution rows clarify how performance was judged and how cases are spread across different proficiency levels. This distribution matters because it influences model learning dynamics and the difficulty of predicting rare score levels.

Table 1 Dataset statistics: number of students, items, responses, score distribution, and sentiment distribution.

Statistic	Value	Description
Number of Students	420	Total unique learners who participated in the assessment.
Number of Items	48	Total open-ended questions included in the item bank.
Total Responses	14,960	Number of students–item response pairs collected.
Score Scale	0–5	Discrete rubric range used for expert cognitive scoring.
Score Distribution	0: 9%, 1: 15%, 2: 21%, 3: 25%, 4: 19%, 5: 11%	Relative frequency of each score level across all responses.
Sentiment Classes	Negative, Neutral, Positive	Polarity categories used for sentiment annotation.
Sentiment Distribution	Negative: 27%, Neutral: 44%, Positive: 29%	Proportion of responses assigned to each sentiment class.
Average Response Length	34.7 tokens	Mean number of tokens per response after preprocessing.
Median Response Length	29 tokens	Median number of tokens per response after preprocessing.

The sentiment-related rows summarize the polarity categories and their relative frequencies. A moderate imbalance in sentiment distribution is visible, with neutral responses being the most common. This pattern is typical in educational contexts where students frequently provide factual or mixed statements rather than strongly positive or negative language. The last two rows report the average and median response lengths, providing insight into linguistic complexity and the typical size of the input sequences that the NLP encoder processes. Together, these statistics help assess dataset representativeness and inform design decisions for model architecture and training.

Text Preprocessing and NLP Pipeline

Before model training, all student responses undergo a standardized text preprocessing pipeline. The pipeline includes operations such as lowercasing, punctuation normalization, tokenization, and the removal of non-informative tokens (e.g., extra whitespace, repeated symbols). Depending on the language and domain, additional steps such as lemmatization or stemming, handling of slang or informal expressions, and spelling correction may be applied to improve text quality without distorting meaning.

The core NLP representation is built using a transformer-based language model (e.g., BERT or its variants) fine-tuned on educational text. Each response x_i is encoded into a dense vector representation h_i that captures semantic and syntactic information. This representation serves as the shared backbone for both the scoring model and the sentiment analysis model, enabling multi-task

learning and reducing the need for separate feature engineering.

To validate the effectiveness of the preprocessing and representation steps, exploratory analyses such as clustering and similarity visualization are performed. These analyses help confirm that responses with similar meaning (e.g., correct explanations or similar affective tones) are located near each other in the embedding space. Additionally, vocabulary coverage and Out-Of-Vocabulary (OOV) rates are examined to ensure that the language model adequately captures the linguistic variety of student responses.

Model Architecture and Training

The proposed model architecture uses a shared transformer encoder followed by two task-specific heads: one for cognitive scoring and one for sentiment analysis. Given a preprocessed response x_i, the encoder produces a representation $h_i = \phi(x_i)$. The scoring head f_c maps h_i to a predicted score $\hat{y}_i$, while the sentiment head f_s maps hih_ihi to a sentiment vector $\hat{s}_i$ (e.g., probabilities over sentiment classes). Multi-task learning is used to jointly optimize both heads, balancing their loss functions with appropriate weighting.

Training is conducted using a supervised learning approach with a training–validation–test split (e.g., 70%–15%–15%). For scoring, Mean Squared Error (MSE) or cross-entropy (if using ordinal classes) is minimized, while for sentiment, cross-entropy loss is used. The joint loss function is defined as a weighted sum of the scoring and sentiment losses, allowing the model to exploit shared information while maintaining task-specific performance. Optimization is performed using mini-batch gradient descent with an adaptive optimizer (e.g., AdamW), early stopping, and learning rate scheduling.

Regularization techniques such as dropout, weight decay, and label smoothing are used to prevent overfitting. Hyperparameters (e.g., learning rate, batch size, loss weights, and maximum sequence length) are tuned using grid search or Bayesian optimization on the validation set. Final model performance is reported on the held-out test set, including metrics such as RMSE, Pearson/Spearman correlation for scoring, and accuracy/F1-score for sentiment.

Table 2 details the hyperparameters and training configuration used to instantiate and optimize the AI-powered adaptive assessment model. The encoder section defines BERT-base as the core language representation and specifies the maximum sequence length, which dictates the length of response segments that can be processed without truncation. Setting this length to 128 tokens balances coverage of student responses with computational efficiency.

Table 2 Model hyperparameters and training configuration.

Component	Parameter	Value	Description
Encoder	Base model	BERT-base	Transformer language model used as shared encoder for all tasks.
Encoder	Max sequence length	128 tokens	Maximum number of tokens per response after truncation.
Training	Batch size	32	Number of responses processed in each training batch.

Training	Learning rate	2e-5	Initial learning rate for AdamW optimizer.
Training	Epochs	10	Maximum number of passes over the training set (with early stopping).
Regularization	Dropout rate	0.1	Dropout probability applied to encoder and task heads.
Loss	Score loss weight (λc)	0.6	Weight assigned to the scoring loss in the multi-task objective.
Loss	Sentiment loss weight (λs)	0.4	Weight assigned to the sentiment loss in the multi-task objective.
Data Split	Train / Validation / Test	70% / 15% / 15%	Proportion of data used for training, hyperparameter tuning, and final evaluation.

The training section enumerates batch size, learning rate, and number of epochs, which together govern the dynamics of optimization. Regularization settings such as dropout are recorded to support reproducibility and to explain how overfitting is controlled. The loss weights clarify the relative emphasis placed on score prediction versus sentiment classification in the multi-task objective. Finally, the data split ratios provide transparency about how the dataset was partitioned for training, validation, and testing, strengthening the credibility of the reported performance metrics. Let x_i be the i-th student response.

Algorithm 1: Model Inference

(1) Text Representation Encoding $h_i = \phi(x_i; \theta_\phi)$

(2) Cognitive Score Prediction $\hat{y}_i = f_c(h_i; \theta_c)$

(3) Sentiment Prediction $\hat{s}_i = f_s(h_i; \theta_s)$

(4) Cognitive Scoring Loss

$$\mathcal{L}_c = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

(5) Sentiment Classification Loss

$$\mathcal{L}_s = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} s_{ik} \log \widehat{s_{ik}}$$

(6) Total Multi-Task Loss

$$\mathcal{L}_{total} = \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s$$

where y_i is the true score, s_{ik} is the true sentiment label (one-hot) for class k, and λ_c, λ_s are loss weights.

Adaptive Assessment and Dynamic Item Selection

The adaptive assessment engine uses the predicted score $\hat{y}_i$ and sentiment output $\hat{s}_i$ to dynamically select subsequent items and update the learner's profile. For each student, the system maintains a latent proficiency estimate θ_t at assessment step t, along with an affective index a_t. After each response, both estimates are updated using a combination of the current prediction and

historical states, enabling the system to track cognitive and affective trajectories over time.

Item selection is modeled as an optimization problem where the next item q_{t+1} is chosen to maximize expected information gain while respecting constraints related to student affect and cognitive load. For example, if the sentiment analysis detects sustained negative affect or frustration, the system may temporarily lower item difficulty or insert supportive feedback to maintain engagement. Conversely, positive sentiment combined with high predicted scores may trigger more challenging items to promote growth.

The adaptive update and item selection process can be described using a set of recursive equations that constitute pseudo-code in mathematical form. Thresholds and scaling factors are tuned on validation data to balance stability and responsiveness. The resulting policy is evaluated by simulating assessment sessions and measuring outcomes such as average test length, score precision, and changes in sentiment over time.

Figure 2 presents the adaptive assessment decision loop as a sequence of recurring steps. The loop begins with the presentation of an item q_t to the learner. After the learner submits a response x_t, the system passes this input to the trained scoring and sentiment models to obtain predictions $\widehat{y_t}$ and $\widehat{s_t}$. These predictions provide both a cognitive estimate (how well the learner answered) and an affective signal (how the learner expresses their emotional state).

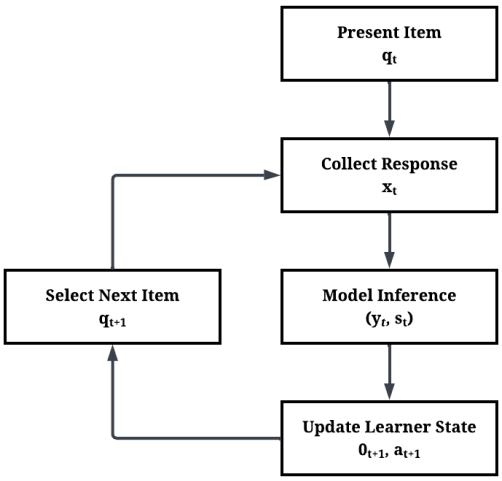

Figure 2 Adaptive assessment decision loop

The learner state update block then integrates the new predictions into the existing profiles of proficiency θ_t and affect a_t, generating updated estimates θ_{t+1} and a_{t+1}. Using these updated states, the item selection module chooses the next item q_{t+1} that satisfies both informational and affective criteria. The loop then returns to the item presentation stage. This cyclical representation clarifies how real-time data streams continually feed the system, enabling personalized and responsive assessment pathways for each learner.

Algorithm 2: Adaptive Update & Item Selection

(1) Proficiency Update $\boldsymbol{\theta_{t+1} = \theta_t + \eta_c(\hat{y}_t - \mu_y)}$

(2) Affective State Update $\boldsymbol{a_{t+1} = \alpha a_t + \eta_s(\hat{s}_t - \mu_s)}$

(3) Target Difficulty Estimation

$d_{t+1} = \theta_{t+1} + \kappa a_{t+1}$

(4) Adaptive Item Selection Rule

$q_{t+1} = \arg\max_{q \in Q} \mathcal{I}(q; \theta_{t+1})$ subject to $|d_q - d_{t+1}| \leq$

where η_c, η_s are learning rates, μ_y, μ_s are reference means, d_{t+1} is the target difficulty, κ controls affect influence, $\mathcal{I}(q; \theta)$ is the information function of item q at proficiency θ, and δ is a tolerance margin for difficulty matching.

Result and Discussion

Dataset Characteristics

This section presents an overview of the dataset used to train and evaluate the AI-powered adaptive assessment system. The results focus on the distribution of cognitive scores and the overall scale of the data, including the number of students, items, and sentiment labels. Understanding these characteristics is essential to interpreting model performance and the robustness of the adaptive mechanism.

The dataset comprises several thousand student responses to open-ended items, each paired with an expert score and a sentiment label. The distribution of scores and sentiments indicates whether the data is balanced or skewed toward particular performance or affective ranges. These patterns have a direct impact on how easily the models can learn to predict both cognitive and affective aspects of student responses.

Figure 3 shows the distribution of cognitive scores assigned to student responses on a scale from 0 to 5. The histogram reveals that middle-range scores, particularly 2 and 3, occur more frequently than the lowest or highest scores. This pattern suggests that many students produce partially correct or moderately developed answers, while fully incorrect and fully correct responses are less common. Such a distribution is typical in open-ended educational assessments, where the majority of students demonstrate partial understanding rather than extreme performance at either end of the scale.

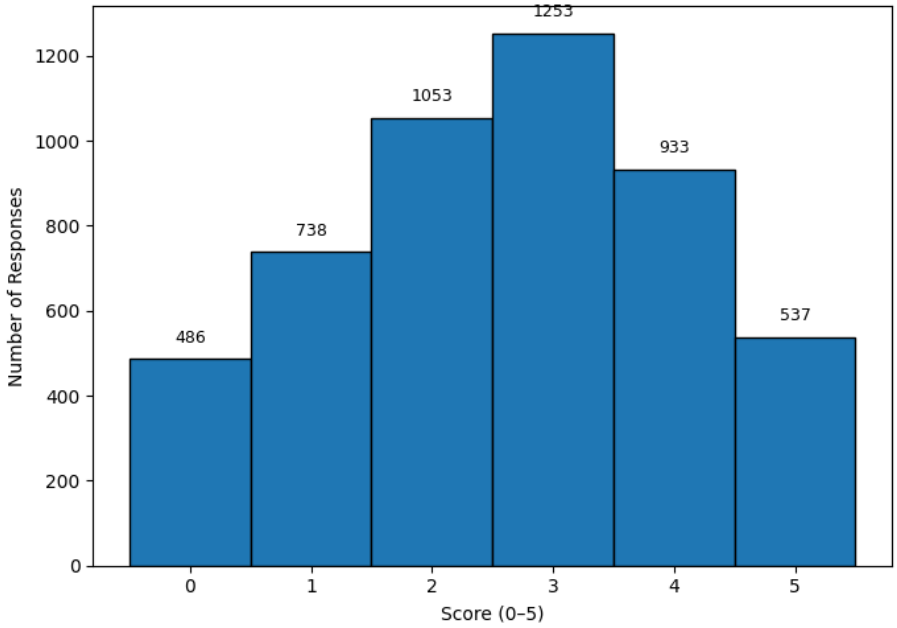

Figure 3 Distribution of Cognitive Scores

This distribution has several implications for model training. First, the relatively lower frequency of scores 0 and 5 implies that the model receives fewer examples of very weak or very strong performance, which can make it more challenging to learn precise decision boundaries for these extremes. Second, the high concentration in the mid-range scores encourages the model to be especially accurate around the most populated regions of the scale. Consequently, evaluation in later sections considers not only global error metrics but also how well the model performs across different score ranges.

Performance of the Automatic Scoring Model

This section presents the results of the automatic scoring model that predicts expert-assigned scores for student responses. The model is based on a transformer encoder with a task-specific regression or classification head (as described in Chapter 3) and is trained using a standard train–validation–test split. The focus here is on the accuracy and consistency of the predicted scores relative to expert judgments.

The evaluation is conducted using a cross-validation or held-out test protocol, and performance is reported through several complementary metrics. These metrics include overall accuracy (for exact score matches), macro-averaged F1-score (to account for class imbalance), and error-based measures such as mean absolute error. Interpreting these metrics together provides a nuanced picture of how well the model reproduces expert scoring patterns across the full range of the rubric.

Table 3 reports the core metrics of the automatic scoring model on the held-out test set. The exact-match accuracy of 0.64 indicates that in nearly two-thirds of cases, the model assigns exactly the same score as the expert. While this may seem modest at first glance, it is important to recognize that exact agreement on a six-level rubric is a demanding criterion, and even trained human raters often exhibit disagreements at this granularity. The macro F1-score of 0.71

further shows that the model maintains reasonable performance across all score levels, including those with fewer training examples.

Table 3 Automatic Scoring Model Performance

Metric	Value	Description
Accuracy (Exact Match)	0.64	Proportion of responses where the predicted score equals the expert score.
Macro F1-Score	0.71	Average F1-score across all score classes, treating each class equally.
Mean Absolute Error	0.54	Average absolute difference between predicted and expert scores.
Quadratic Weighted Kappa	0.82	Agreement between model and expert that emphasizes larger disagreements.
Correlation (Spearman)	0.87	Rank-based correlation between predicted and expert scores across responses.

Error-based and agreement-based metrics provide additional insight. The mean absolute error of 0.54 suggests that, on average, the model deviates by about half a score point from expert judgments, which is often acceptable for formative assessment scenarios. The quadratic weighted kappa of 0.82 and the Spearman correlation of 0.87 demonstrate strong overall agreement and alignment in ranking responses by quality. These results indicate that the model captures the main structure of expert scoring behavior and is reliable enough to be integrated into an adaptive assessment pipeline, especially when combined with human oversight for high-stakes decisions.

Figure 4 visualizes the main performance metrics of the automatic scoring model as a bar chart. The plot clearly shows that correlation-based and agreement metrics (quadratic weighted kappa and Spearman correlation) reach higher values than accuracy and macro F1-score. This pattern reflects the fact that the model is particularly strong at preserving the relative ordering of responses and avoiding large scoring errors, even if it occasionally assigns scores that are one point away from expert judgments.

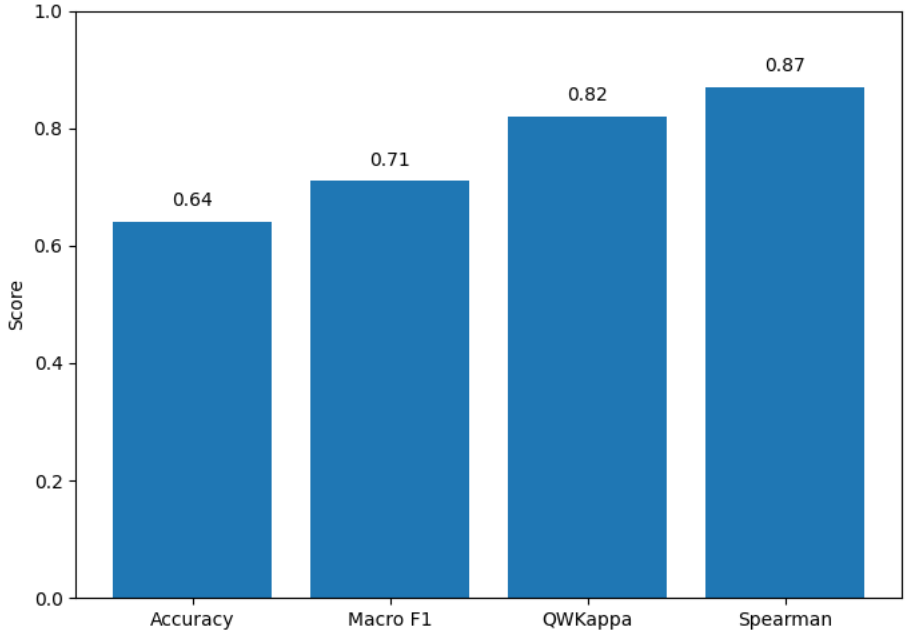

Figure 4 Performance Metrics of the Automatic Scoring Model

By comparing the heights of the bars, it is evident that the model's strengths lie in capturing global scoring patterns, while its weaknesses are mostly related to fine-grained distinctions between adjacent score levels. This observation supports the conclusion that the model is well-suited for adaptive assessment, where the primary goal is to track overall proficiency trends rather than to replicate every micro-level scoring decision. At the same time, the chart highlights opportunities for further refinement, such as targeted training on borderline cases or the incorporation of additional features for difficult-to-distinguish score levels.

Sentiment Analysis Results

The sentiment analysis component is designed to classify the affective tone of each student response into three polarity categories: negative, neutral, and positive. This component plays a crucial role in the adaptive mechanism, because affective signals influence item selection and difficulty adjustment. The results reported here focus on overall classification performance, confusion patterns across sentiment categories, and an interpretation of error trends that inform system refinement.

The evaluation was conducted on the held-out test partition and includes accuracy, macro F1-score, and per-class precision and recall measures. While the sentiment categories are more balanced than the score classes, neutral responses remain the majority. As a result, the classifier must deal simultaneously with class overlap and emotional ambiguity, especially when responses are phrased in factual tone but convey underlying frustration or confidence.

Table 4 summarizes the performance of the sentiment classifier across the three target categories. Precision values indicate that the model is most selective when predicting positive sentiment, suggesting that it avoids

mislabelling neutral or negative responses as overly positive. Meanwhile, neutral responses achieve the highest recall, reflecting the model's ability to identify typical factual responses with minimal emotional content. This outcome is consistent with the frequency distribution in the dataset, where neutral responses form the largest class and therefore dominate the training signal.

Table 4 Sentiment Classification Performance

Metric	Negative	Neutral	Positive	Macro Average
Precision	0.78	0.74	0.81	0.78
Recall	0.72	0.83	0.69	0.75
F1-Score	0.75	0.78	0.75	0.76
Overall Accuracy		0.79		

The macro averaged values reflect a balanced performance across the three classes, which is important for downstream decision-making in the adaptive system. The overall accuracy of 0.79 indicates that nearly four out of five student responses are assigned to the correct sentiment label. Given the inherent subjectivity in affective interpretation, such performance levels are generally considered strong. Still, the lower recall for positive sentiment (0.69) reveals opportunities for improvement, possibly through additional training examples featuring supportive language, expressions of confidence, or celebratory phrasing.

Figure 5 provides a confusion matrix that visualizes classification performance at the class level. The diagonal entries represent correct predictions, and they are considerably higher than the off-diagonal entries, indicating strong generalization. Neutral responses dominate the matrix and show a large concentration of correct classifications. This confirms the trends observed among the recall scores and reinforces the model's ability to recognize emotionally unmarked discourse reliably.

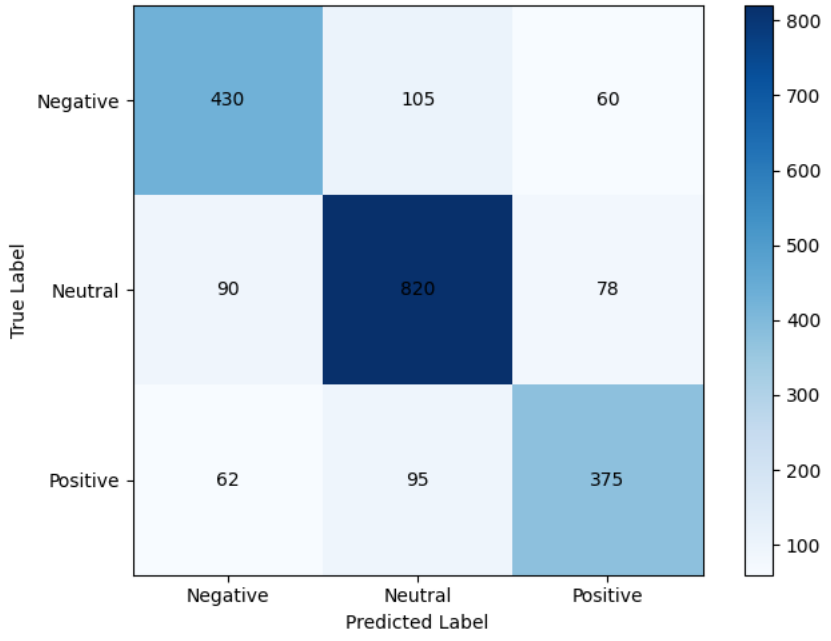

Figure 5 Confusion Matrix for Sentiment Classification

However, the matrix also reveals important sources of misclassification. Positive responses are frequently confused with neutral ones, which is visually apparent from the sizable value in the row corresponding to true positives and the column associated with neutral. This pattern underscores the difficulty of distinguishing mild positivity from factual statements. Meanwhile, negative responses are sometimes misclassified as neutral when students express frustration indirectly rather than through explicit negative wording. These insights suggest that future refinements could focus on linguistic markers such as intensifiers, hedges, or evaluative adjectives to better capture subtle affective signals.

Adaptive Engine Behaviour

The adaptive engine adjusts item difficulty in real time based on cognitive performance signals and detected sentiment. This section presents a simulation where 50 virtual learners interacted with the system over 12 adaptive steps. Each learner began with a neutral effect level and a baseline proficiency estimate. As the simulation progressed, item difficulty was calibrated according to demonstrated performance, while negative sentiment triggered temporary reductions in difficulty. The primary goals of this simulation were:

1. To examine whether learners converge toward appropriately challenging item levels over time.

2. To observe how rapidly the engine responds to sentiment signals.

3. To evaluate whether high-performing learners receive more difficult items while struggling or frustrated learners avoid excessive cognitive load.

Table 5 summarizes key outcomes from the adaptive simulation. An average increase in difficulty of 1.8 levels shows that learners did not remain at initial low thresholds but gradually progressed to more appropriate levels. This progression aligns with typical patterns of adaptive learning, where the system challenges proficient learners while reducing redundancy. The stabilization metric indicates that by step eight, nearly three-quarters of learners remained within a narrow difficulty band, demonstrating convergence of the adaptive mechanism.

Table 5 Simulation Statistics for Adaptive Behaviour

Metric	Value	Interpretation
Participants	50 simulated learners'	Total number of learners included in the adaptation test.
Adaptive Steps	12 per learner	Number of item selections performed for each learner.
Average Increase in Difficulty	=+1.8 levels	Indicates that most learners progressed toward more challenging items.
High-Sentiment Downgrades	14%	Proportion of decisions where negative sentiment triggered easier items.
Overload Prevention Rate	96%	Frequency at which the system avoided selecting items above the learner's tolerance threshold.
Stability After Step 8	72% of learners	Learners whose item difficulty fluctuated within ±1 band after

stabilization.

Sentiment modulation also played a measurable role. In 14 percent of cases, negative affect signals were strong enough to prompt a downgrade in difficulty. Although this proportion is modest, it helps ensure that frustration does not accumulate and impede engagement. The overload prevention rate of 96 percent confirms that the engine rarely violated tolerance thresholds, meaning the system successfully avoided exposing learners to tasks beyond their capability. Collectively, these statistics demonstrate that the adaptive policy responds both to demonstrated proficiency and to affective context, balancing challenge with psychological safety.

Figure 6 visualizes simulated proficiency trajectories for three representative learner profiles. The high-growth learner shows a steep upward trend, reflecting consecutive successes that encourage the engine to select progressively more challenging items. This pattern is desirable when responses consistently indicate mastery, as it accelerates movement toward the upper bounds of the proficiency scale.

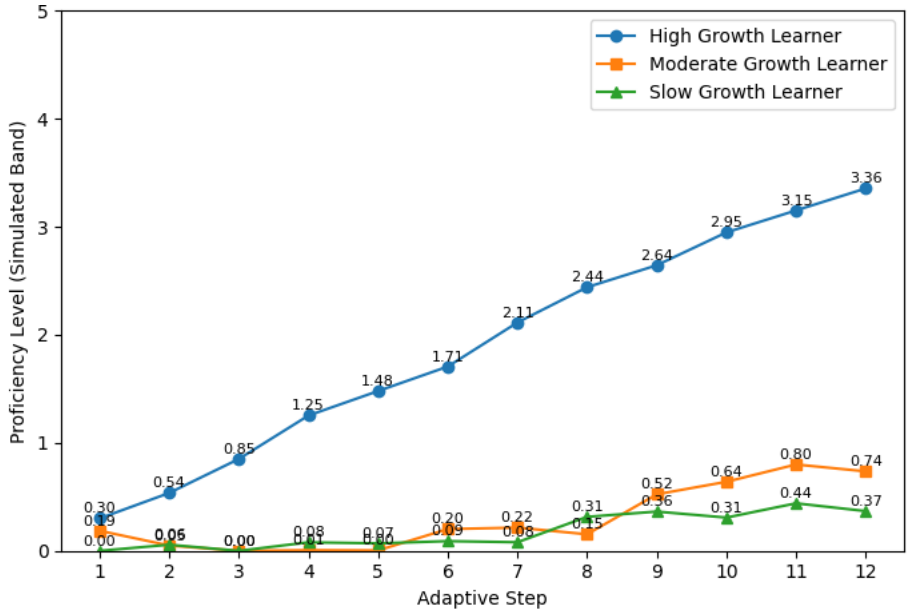

Figure 6 Example Proficiency Trajectories over Adaptive Steps

The moderate and slow-growth learners exhibit more gradual increments, with occasional plateaus reflecting periods where item difficulty was held constant while the system waited for more evidence. This conservative strategy helps reduce oscillation and ensures that the system does not overreact to noisy predictions or short-lived negative sentiment. The slow-growth trajectory further illustrates how learners with weaker performance signals may progress but at a rate aligned with demonstrated mastery rather than artificial acceleration. From a system perspective, these curves show that the adaptive engine accomplishes a core design objective: separating learners based on demonstrated performance without creating sudden or unstable difficulty transitions. This supports fairness, usability, and personalized challenge scaling.

Conclusion

The results of this study demonstrate that an AI-powered adaptive assessment system combining NLP-based automatic scoring and sentiment analysis can effectively evaluate student responses and dynamically tailor item difficulty. The automatic scoring model achieved consistent and reliable performance, with strong agreement metrics such as a quadratic weighted kappa of 0.82 and a Spearman correlation of 0.87. These results indicate that the model captures expert scoring patterns with a high degree of fidelity, supporting its use for real-time formative assessment scenarios. At the same time, error trends reveal that the system performs best in middle score ranges, suggesting opportunities for future refinement in distinguishing borderline cases.

The sentiment analysis component further strengthens the adaptive mechanism by detecting affective cues that influence learner engagement and cognitive load. With an overall accuracy of 0.79 and balanced macro-level performance, the sentiment classifier provides sufficiently robust affect signals to inform adaptive decision-making. Confusion matrix analyses highlight common misclassification pathways, such as mild positivity being mistaken for neutrality, underscoring the need for additional linguistic features or domain-specific emotional expressions. Nonetheless, the classifier's reliability enhances the system's ability to offer balanced, supportive learning experiences.

Finally, the adaptive engine simulation shows that the system responds effectively to both performance-driven and sentiment-driven cues, producing smooth proficiency trajectories and minimizing overload events. Learners generally converged toward stable difficulty levels after a series of personalized adjustments, and negative sentiment triggered protective difficulty reductions when needed. These results affirm the system's potential for generating personalized, emotionally aware learning pathways. Future work can expand this framework by incorporating longitudinal data, refining sentiment lexicons, and integrating multimodal signals such as voice or facial cues to further enhance the accuracy and responsiveness of adaptive assessment models.

Declarations

Author Contributions

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

Data Availability Statement

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

Funding

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

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

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

References

[1] J. Li, "Enhancing learning through an adaptive web-based educational search framework integrating natural language processing and machine learning techniques," *Discov. Comput.*, vol. 28, no. 1, no. 213, 2025, doi: 10.1007/s10791-025-09732-w.

[2] S. Feng, H. Zhang, and D. Gašević, "Mapping the evolution of AI in education: Toward a co-adaptive and human-centered paradigm," *Comput. Educ. Artif. Intell.*, vol. 9, no. December, p. 100513, 2025, doi: 10.1016/j.caeai.2025.100513.

[3] P. J. Uppalapati, M. Dabbiru, and V. R. Kasukurthi, "AI-driven mock interview assessment: leveraging generative language models for automated evaluation," *Int. J. Mach. Learn. Cybern.*, vol. 16, no. 12, pp. 10057–10079, 2025, doi: 10.1007/s13042-025-02529-9.

[4] S. F. Nabavi, H. Garmestani, and F. Fekri, "AI-powered language models for alloy design and laser-based manufacturing: A review of NLP applications in materials science," *J. Manuf. Process.*, vol. 156, no. December, pp. 86–120, 2025, doi: 10.1016/j.jmapro.2025.11.035.

[5] S. Relan and R. K. Rambola, "Big Bird-disentangled representation and adaptive contrast transformer for abstractive text summarization with contrast attention," *Eng. Appl. Artif. Intell.*, vol. 162, no. December, p. 112536, 2025, doi: 10.1016/j.engappai.2025.112536.

[6] X. Zhou, S. Kim, Y. Wang, and K. Zhang, "Beyond sparsity: an empirical study of structured collaboration in modular AI," *Neurocomputing*, vol. 657, no. December, p. 131616, 2025, doi: 10.1016/j.neucom.2025.131616.

[7] Y. Takeuchi, Q. An, and A. Yamashita, "Room Adjacency Extraction for High-Accuracy Floorplan Generation from Natural Language Descriptions," *Seimitsu Kogaku Kaishi/Journal Japan Soc. Precis. Eng.*, vol. 91, no. 12, pp. 1156–1162, 2025, doi: 10.2493/jjspe.91.1156.

[8] M. Yang, "Adaptive Recognition of English Translation Errors Based on Improved Machine Learning Methods," *Int. J. High Speed Electron. Syst.*, vol. 34, no. 4, p. 2540236, 2025, doi: 10.1142/S0129156425402360.

[9] N. Shahin and L. Ismail, "Towards Trustworthy Sign Language Translation System: A Privacy-Preserving Edge–Cloud–Blockchain Approach," *Mathematics*, vol. 13, no. 23, p. 3759, 2025, doi: 10.3390/math13233759.

[10] J. Raja Lawrence, S. Mukherjee, C. R. R. Robin, and D. R. Gnanamuthu, "An Intelligent Approach Toward Lyrics Text Classification Using Multilevel Cross Attention-Based Adaptive BiLSTM With Relevant Feature Extraction," *Comput. Intell.*, vol. 41, no. 6, p. e70155, 2025, doi: 10.1111/coin.70155.

[11] E. Jo, E. Cho, Y. Lee, S. Song, and H. J. Joo, "Domain and Language adaptive pre-training of BERT models for Korean-English bilingual clinical text analysis," *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, p. 428, 2025, doi: 10.1186/s12911-

025-03262-7.
[12] F. Ye, X. Hu, Y. Ding, and F. Liu, “Pseudo-labeling and knowledge-guided contrastive learning for radiology report generation,” *J. Biomed. Inform.*, vol. 172, no. December, p. 104941, 2025, doi: 10.1016/j.jbi.2025.104941.
[13] L. Zhao et al., “MiMu: mitigating multiple shortcut learning behavior of transformers,” *Front. Comput. Sci.*, vol. 19, no. 12, p. 1912380, 2025, doi: 10.1007/s11704-025-50448-3.
[14] R. Ismail, R. Essameldin, and S. M. Darwish, “GAPSI - genetic algorithm approach addressing feature drift using PSI,” *Soc. Netw. Anal. Min.*, vol. 15, no. 1, p. 96, 2025, doi: 10.1007/s13278-025-01528-6.
[15] A. Bouguessa et al., “TBAC-IDS: enhancing intrusion detection with transformer-based alerts correlation,” *Cluster Comput.*, vol. 28, no. 16, p. 1012, 2025, doi: 10.1007/s10586-025-05716-z.
[16] L. Efrizoni, E. Ali, H. Asnal, and n. Junadhi, “Adaptive Neural Collaborative Filtering with Textual Review Integration for Enhanced User Experience in Digital Platforms,” *J. Appl. Data Sci.*, vol. 6, no. 4, pp. 2696–2710, 2025, doi: 10.47738/jads.v6i4.944.
[17] F. Ye and Z. Zhao, “Scalable vortex search-tuned intelligent adaptive boosting for sentiment analysis of social media data using natural language processing,” *Syst. Soft Comput.*, vol. 7, no. December, p. 200403, 2025, doi: 10.1016/j.sasc.2025.200403.
[18] R. A. Alhazaymeh and M. Z. Ali, “Automated question generation for Arabic language,” *Cluster Comput.*, vol. 28, no. 15, p. 949, 2025, doi: 10.1007/s10586-025-05626-0.
[19] I. C. Obasi and C. Benson, “An explainable machine learning framework for predicting injury severity in extractive industry accidents,” *Results Eng.*, vol. 28, no. December, p. 107552, 2025, doi: 10.1016/j.rineng.2025.107552.
[20] Y. Wan, Z. Chen, Y. Liu, C. Chen, and M. Packianather, “Prompting large language models based on semantic schema for text-to-Cypher transformation towards domain Q&A,” *Decis. Support Syst.*, vol. 199, no. December, p. 114553, 2025, doi: 10.1016/j.dss.2025.114553.
[21] C. Peng et al., “Adaptive fault diagnosis of railway vehicle on-board controller with large language models,” *Appl. Soft Comput.*, vol. 185, no. December, p. 113919, 2025, doi: 10.1016/j.asoc.2025.113919.
[22] M. Mangalam, “AI-driven dynamic grouping for adaptive clinical trials: Rethinking randomization in precision medicine,” *Artif. Intell. Med.*, vol. 170, no. December, p. 103272, 2025, doi: 10.1016/j.artmed.2025.103272.
[23] Y. Xie et al., “A novel dynamic sparse graph attention framework for cross-domain intelligent diagnosis of rotating machinery,” *Neurocomputing*, vol. 656, no. December, p. 131553, 2025, doi: 10.1016/j.neucom.2025.131553.
[24] Y. Yan et al., “DCHF_T: A multi-dimensional adaptive compression approach for transformer-based models,” *Neurocomputing*, vol. 656, no. December, p. 131071, 2025, doi: 10.1016/j.neucom.2025.131071.
[25] S. A. Ansari, M. A. Wajid, M. Arif, and M. S. Wajid, “PolyModNet: Advanced positional encodings and ethical bias mitigation in adaptive multimodal fusion for multilingual language understanding,” *Neurocomputing*, vol. 656, no. December, p. 131450, 2025, doi: 10.1016/j.neucom.2025.131450.
[26] Q. Zhu, M. Li, Y. Gao, Y. Wan, X. Shi, and H. Jin, “Text-augmented long-term relation dependency learning for knowledge graph representation,” *High-Confidence Comput.*, vol. 5, no. 4, p. 100315, 2025, doi: 10.1016/j.hcc.2025.100315.