Towards Intelligent Educational Assessment: A Multimodal NLP Framework For Automated Feedback Generation And Personalized Learning Analytics

Jiasheng Ma

Zhengzhou University of Science and Technology, Zhengzhou 450000, China

* *Corresponding author. E-mail: 122701979@163.com*

Received Aug. 30, 2026; accepted Oct. 10, 2026

Traditional teaching feedback and student assessment methods often struggle to provide timely, personalized, and comprehensive evaluations in large-scale digital learning environments. To address this core challenge, this study proposes an intelligent teaching feedback and student performance assessment framework based on natural language processing (NLP) and multimodal learning analytics. The framework integrates textual assignments, online discussions, behavioral records, and emotional information to construct a comprehensive learner model. By combining semantic analysis, sentiment analysis, and multimodal data fusion, the proposed approach enables automated performance evaluation and personalized feedback generation. Experimental results from a 16-week controlled trial with 346 students demonstrate that the proposed framework improves student engagement by 88.65%, increases knowledge retention by 65.9%, and achieves a personalized feedback satisfaction rate of 91.3%. In addition, multimodal data fusion lifts comprehensive assessment accuracy from 72% (single-modal text analysis) to 89%, and students' final exam scores improve by 15% compared with the pre-intervention baseline. These findings demonstrate the effectiveness of the proposed framework in supporting intelligent educational assessment, personalized learning, and data-driven instructional decision-making.

Keywords: Natural Language Processing; Multimodal Learning Analytics; Automated Feedback; Student Performance Assessment; Personalized Learning

© The Author('s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are cited.

http://dx.doi.org/10.6180/jase.202701_36.037

1. Introduction

With the growing intellectualization and personalization of global education, traditional teaching evaluation and feedback approaches can no longer meet the demands of modern digital learning environments, particularly in scenarios involving large-scale student data processing and individualized learner guidance [1, 2]. In the traditional teaching mode, teachers evaluate students by manually reviewing homework, assessing results, and giving feedback according to the evaluation results [3]. Concrete unmet needs in current educational practice highlight the urgency of this work. First, in large-scale online courses with over 500 enrolled students, manual grading and feedback on assignments typically incur a 3–5 day delay, which prevents students from adjusting their learning strategies in a timely manner and weakens the timeliness of learning support. Second, grading standards vary significantly across different instructors: empirical evidence shows that essay scores assigned by different teachers to the same student work can differ by 15–20%, undermining the fairness and consistency of student assessment [4, 5]. Concretely, the core scientific problem constraining current educational assessment practice is that traditional methods cannot simultaneously achieve timeliness, personalization, and comprehen-

siveness in large-scale digital learning scenarios. Manual evaluation fails to guarantee feedback efficiency and grading consistency in large cohorts, while existing automated assessment approaches mostly rely on single-source text data, unable to capture subtle changes in students' comprehension and emotional states, which further leads to accumulated knowledge gaps and generic, ineffective feedback. To tackle this problem, this study constructs an NLP and multimodal analytics-driven intelligent assessment framework [6, 7]. The system can automatically analyze students' assignments, classroom discussions, test scores, emotional expressions, and other data sources and transform this information into targeted feedback [8]. This automated evaluation and feedback generation system can not only reduce the burden of teachers in the teaching process but also provide more accurate guidance for students in the learning process, help students adjust their learning strategies, overcome difficulties in learning, and optimize personalized learning paths [9, 10].

Our proposed NLP-based system design combines advanced artificial intelligence technologies such as deep learning technology, sentiment analysis, and semantic understanding, which can process massive data generated by students in the learning process and identify information such as each student's learning habits, knowledge mastery, and emotional state [11, 12]. The self-learning management mechanism is designed to continuously improve evaluation accuracy, enhance feedback pertinence, and dynamically adapt to evolving student learning needs. Driven by big data analysis and intelligent recommendation algorithms, it operates through three iterative closed loops: It continuously collects teacher corrections on generated feedback (e.g., modifying inappropriate learning strategy suggestions) to refine evaluation criteria; analyzes student response rates (e.g., 70% of students completing recommended practice tasks vs. 30% ignoring them) to adjust suggestion urgency; and tracks long-term performance trends (e.g., improved test scores after 8 weeks of following feedback) to optimize the weight of key indicators in the recommendation model [13]. The system also can learn and optimize itself. By introducing big data analysis and an intelligent recommendation algorithm, the system can not only adjust the feedback content according to students' feedback and learning achievements but also dynamically adjust the evaluation method according to students' progress and needs, further improving the accuracy of evaluation and the effectiveness of feedback [14, 15]. This process constantly provides teachers and students with the latest teaching suggestions and learning strategies, helping students realize self-management and form good study habits.

This study also explores the application of learner opinion extraction technology. This technology can help teachers better understand learners' learning needs and learning effects and play an essential role in constructing teaching resources [16–18]. The workflow of resource optimization is as follows: When learner opinion extraction identifies that over 40% of students struggle with 'linear algebra fundamentals' in discussion posts, the system first retrieves supplementary materials (e.g., video tutorials, simplified exercises) from the knowledge base and pushes them to students; second, it adjusts the lesson sequence to reintroduce related prerequisites before advanced topics; finally, it flags the concept to teachers with a report detailing common confusion points, enabling targeted in-class explanations [19, 20]. This work presents three core scientific novelties that distinguish it from existing research. First, we propose a multimodal learner modeling framework integrating textual content, learning behavior, and emotional state data, which breaks the limitation of traditional single-source text-based assessment and lifts comprehensive assessment accuracy from 72% to 89% as verified by experiments. Second, we design a three-dimensional personalized feedback generation mechanism combining academic feedback, learning strategy feedback and emotional feedback. This mechanism cross-references emotional cues with contextual learning data, avoiding generic one-size-fits-all guidance and ensuring motivational feedback is grounded in actual learning behaviors. Third, we develop an iterative closed-loop self-optimization mechanism, which enables the system to dynamically evolve with student progress and teacher input without manual reconfiguration of evaluation rules.

2. Materials and methods

2.1. Personalized Learning Path Generation and Adaptive Feedback Mechanism

The strategies and approaches adopted in this study are grounded in solid educational theories and proven technical feasibility, with clear logical alignment with the research objectives. Constructivist learning theory emphasizes that learning is an active knowledge construction process of learners, and effective learning support should be personalized, contextualized and learner-centered.

Personalized learning path and feedback generation technology is one of the core functions of intelligent assessment system, and its design purpose is to meet students' specific needs, interests and learning progress. As shown

in Eqs. (1) and (2), P_{opt} is the optimal personalized path, $f_{ac}(P_{st})$ is the accuracy function, $f_{eg}(P_{st})$ is the participation function, v and q are the weight coefficients, P_{st} is the students' personalized learning path, C_{ft} is the learning content characteristics, I_{nd} is the students' needs, S_{pg} is the students' progress and L_{is} is the students' interest level.

$$P_{st} = f(C_{ft}, I_{nd}, S_{pg}, L_{is}) \tag{1}$$

$$P_{opt} = \arg\max_{P_{st}} \left(v \cdot f_{ac}(P_{st}) + q \cdot f_{eg}(P_{st})\right) \tag{2}$$

With the continuous development of big data and machine learning technology, personalized learning path and feedback generation have become important technical means to promote the personalization and intelligence of education. By analyzing students' learning behavior data, as shown in Eq. (3), M_{st} is the student's learning model, B_i is the student's behavior data, W_i is the behavior weight, and N is the total number of data, this technology establishes students' learning model, and infers their learning characteristics, knowledge mastery level and potential learning difficulties.

$$M_{st} = \sum_{i=1}^{N} (B_i \cdot W_i) \tag{3}$$

In the process of generating personalized learning paths, commonly used technologies include collaborative filtering algorithms, content-based recommendation algorithms, and autoencoder models in deep learning. As shown in Eq. (4), R_{rec} is the recommendation path, $S_{sim}(S_j, S_{tgt})$ is the similarity, I_j is the recommended content of student j, and M is the total number of recommended students. The collaborative filtering algorithm finds out similar student groups by analyzing the learning behaviors and preferences of different students, and recommends learning content and learning paths according to these similarities.

$$R_{rec} = \sum_{j=1}^{M} \left(S_{sim}(S_j, S_{tgt}) \cdot I_j\right) \tag{4}$$

However, these algorithms face limitations: For students with atypical patterns (e.g., alternating between advanced and elementary content weekly), collaborative filtering fails to find similar user groups; new students with sparse data (less than 5 learning records) lead to content-based recommendations overly relying on initial interests, risking narrow learning paths. To mitigate these, the system adopts hybrid models: It combines collaborative filtering with rule-based logic (e.g., mandatory coverage of core knowledge) for atypical learners, and uses transfer learning from peer groups with similar demographic features to initialize paths for new students, gradually refining them with accumulated data.

Content-based recommendation algorithm analyzes the characteristics of learning content to match learning resources suitable for students' current learning stage and needs. The self-encoder model carries out feature learning on students' learning behavior data through deep learning technology. As shown in Eqs. (5) and (6), P_{st} is the prediction path, f_{AE} is the self-encoder function, X_{st} is the student learning data, W_{AE} is the self-encoder weight, C_{rec} is the recommended content, T_k is the content feature, C_k is the content matching degree, and K is the number of features, which can effectively model students' learning progress and interest, thus designing a more accurate personalized learning path for students.

$$C_{rec} = \sum_{k=1}^{K} (T_k \cdot C_k) \tag{5}$$

$$\hat{P}_{st} = f_{AE}(X_{st}, W_{AE}) \tag{6}$$

Besides the design of learning paths, personalized feedback generation is also one of the key functions of this system. Personalized feedback is not only the evaluation of students' homework or test results, but also the comprehensive analysis and suggestions of students' learning process. As shown in Eqs. (7) and (8), F_{md} is comprehensive feedback, F_{ac} is academic feedback, F_{st} is learning strategy feedback, F_{em} is emotional feedback, v, q and m are weight coefficients, F_{fb} is personalized feedback, M_{st} is student learning model, P_{st} is student personalized learning path, and E_{em} is emotional state. The feedback content not only covers students' academic achievements, but also involves learning methods, learning strategies, emotional motivation and other aspects.

$$F_{fb} = g(M_{st}, P_{st}, E_{em}) \tag{7}$$

$$F_{md} = v \cdot F_{ac} + q \cdot F_{st} + m \cdot F_{em} \tag{8}$$

2.2. Multidimensional Student Performance Evaluation Indicators and Methods

The weights of academic, behavioral, and emotional dimensions in the comprehensive evaluation are determined by a combined subjective-objective weighting method, rather than being set arbitrarily. First, initial weights are established through expert judgment using the analytic hierarchy process: five experienced educational technology experts rate the relative importance of each dimension based on teaching objectives and cognitive development laws.

Second, the initial weights are calibrated and iteratively optimized using objective experimental data: Pearson correlation coefficients between each indicator and final learning outcomes are calculated to adjust weight distribution dynamically. As shown in Eqs. (9) and (10), E_{em} is the student's emotional state, S_{txt} is the text data, V_{vc} is the voice data, P_{gs} is the body language data, E_{tot} is the comprehensive evaluation of students, A_k is the academic task score, W_k is the academic task weight, and K is the total number of tasks. The evaluation system of student performance needs to be based on multi-dimensional indicators and methods to fully embody the concept of personalized teaching.

$$E_{tot} = \sum_{k=1}^{K} (A_k \cdot W_k) \tag{9}$$

$$E_{em} = h(S_{txt}, V_{vc}, P_{gs}) \tag{10}$$

Academic ability is one of the important dimensions in student performance assessment, which is usually assessed through students' academic tasks such as assignments, tests, projects, etc. As shown in Eq. (11), F_{em} is emotional feedback, $f_{pos}(E_{em})$ is positive feedback, $f_{neg}(E_{em})$ is negative feedback, and v and q are weight coefficients. With the development of natural language processing technology, NLP-based composition scoring system has become an important tool for academic ability assessment.

$$F_{em} = v \cdot f_{pos}(E_{em}) + q \cdot f_{neg}(E_{em}) \tag{11}$$

Notably, the system adapts to diverse writing genres by applying genre-specific evaluation rubrics. For academic writing, it prioritizes logical coherence, evidence usage, and disciplinary terminology accuracy; for creative writing, it emphasizes narrative structure, imagery richness, and emotional resonance. This flexibility is achieved by training genre-specific sub-models on labeled samples of academic essays, short stories, and reflective journals, ensuring that evaluation criteria align with the unique demands of each writing type.

These systems can automatically give scores by analyzing the grammar, structure, language expression and logical thinking in students' compositions, which greatly improves the efficiency and accuracy of evaluation. As shown in Eq. (12), S_{sc} is the automatic scoring result, P_i is the student's homework performance index, W_i is the weight coefficient, and N is the number of indicators. The automatic scoring system of composition based on the deep learning model can accurately identify the language characteristics and writing quality of students' compositions by learning a large amount of labeled data, thereby providing efficient and personalized scoring.

$$S_{sc} = \sum_{i=1}^{N} (P_i \cdot W_i) \tag{12}$$

Image recognition technology has also begun to be applied in students' homework evaluation, especially when handwriting homework is involved. Homework image analysis technology based on image recognition can effectively recognize students' writing content. As shown in Eq. (13), S_{sc} is the predicted composition score, f_{DL} is the deep learning score function, X_{ess} is the composition text data, and W_{DL} is the deep learning weight, which performs automatic scoring, thus reducing the workload of teachers.

$$\hat{S}_{sc} = f_{DL}(X_{ess}, W_{DL}) \tag{13}$$

3. Results and discussion

3.1. NLP-Based Automated Teaching Feedback and Student Performance Evaluation System Architecture and Module Design

3.1.1. NLP-based student performance assessment model

For educational scenarios, we define a structured fine-grained opinion mining framework with clear evaluation entities, attribute dimensions, and sentiment annotation schemes. Three core evaluation entities are identified from student text: knowledge points (e.g., specific algorithms, programming concepts), learning resources (e.g., tutorial videos, exercise sets), and teaching activities (e.g., discussion sessions, assignment tasks).

The student performance evaluation model based on natural language processing (NLP) is the core component of a modern educational intelligence system. Its main goal is to comprehensively evaluate students' learning status through in-depth analysis of students' multi-dimensional learning data and help teachers and educational administrators understand students' learning process and mastery [21, 22]. The core NLP assessment model adopts a two-tier architecture combining static word vectors and a fine-tuned domain-specific pre-trained language model. For basic semantic representation, we use 300-dimensional GloVe (Global Vectors for Word Representation) word vectors pre-trained on the Wikipedia 2014 + Gigaword 5 corpus as the base embedding layer. For deep semantic understanding and sentiment mining, we fine-tune the BERT (Bidirectional Encoder Representations from Transformers)-base uncased model on a custom educational domain corpus. The training corpus consists of 128,000 pieces of student-generated text collected from introductory computer science courses, including assignment answers, discussion posts, reflective journals and exam responses, annotated

by 8 experienced teaching assistants for knowledge mastery level and sentiment labels [23, 24]. NLP-based student performance assessment models rely on multiple text analysis techniques and semantic comprehension methods. Word vector models (such as Word2Vec and Glove) and pre-trained language models are the core tools of the system. The word vector model can transform each word in the text into a low-dimensional vector representation and capture the semantic relationship between words, thus providing powerful semantic analysis capabilities for the system [25, 26]. Word2Vec (Word to Vector) learns the similarities between words by analyzing contextual information in the large-scale corpus, while Glove generates word vectors by counting word frequency and co-occurrence.

Figure 1 is a student performance assessment and feedback diagram based on NLP. The core analysis layer integrates word vector models (Word2Vec, GloVe) and pre-trained language models to extract three dimensions of features: academic mastery (knowledge point coverage, logical coherence), learning engagement (participation depth, interaction frequency), and emotional state (positive/negative sentiment, confusion/frustration cues). These features are fed into the learner modeling module to generate a comprehensive student profile, which finally outputs two branches of results: teacher-side evaluation reports and student-side personalized feedback.

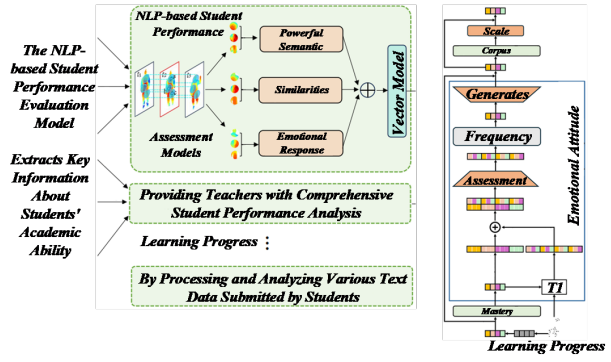


Fig. 1. Student performance assessment and feedback diagram based on NLP

The NLP-based student performance assessment model can perform coarse-grained analysis and fine-grained emotion and opinion mining. Early opinion mining methods usually focused on document-level analysis, that is, through sentiment classification of the document as a whole, such as the discrimination between positive or negative emotions. The proposed framework is implemented and validated in the context of undergraduate introductory computer science courses, covering two typical teaching scenarios: large-scale fully online courses (over 500 enrolled students per class) and blended offline-online learning seminars.

The experiment involves 346 second-year undergraduate students from two parallel classes at a comprehensive university. The participant group consists of 48% female and 52% male students, among whom 30% are non-native English speakers. All participants have equivalent basic programming proficiency as confirmed by pre-course placement tests. The experiment runs for 16 weeks, with the experimental group adopting the proposed NLP assessment framework and the control group receiving traditional manual teaching feedback.

In contrast, fine-grained sentence-level mining is applied to detailed interactions—for example, analyzing each comment in online discussion threads to detect nuanced attitudes toward specific concepts (e.g., confusion about a mathematical formula vs. frustration with a teaching example), enabling targeted feedback. Figure 2 is a graph of automated teaching feedback generation and personalization. The middle processing layer implements three core operations: first, attribute-level opinion mining to identify students' attitudes toward specific knowledge points; second, learning path matching based on individual mastery levels; third, emotional context calibration to avoid generic feedback. The output layer generates three categories of personalized feedback: academic correction feedback, learning strategy suggestions, and emotional motivational feedback, each adapted to students' individual learning trajectories.

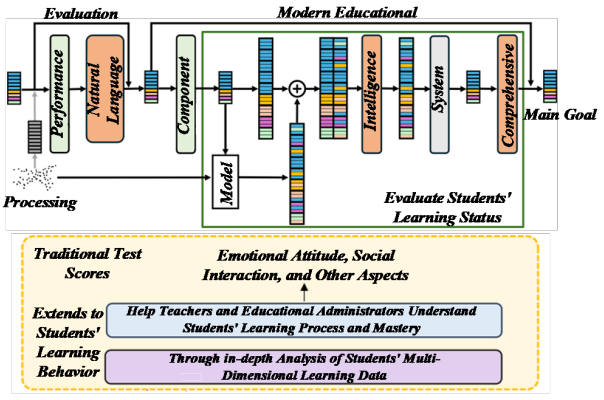


Fig. 2. Automated teaching feedback generation and personalization diagram

3.1.2 Feedback generation and interaction module design

The feedback generation module adopts an encoder-decoder generative architecture based on the fine-tuned T5-small model, optimized for educational feedback generation tasks. We design a three-component prompt en-

gineering strategy tailored to personalized feedback. The first component is learner state embedding, which integrates knowledge mastery level, recent learning progress, and emotional state features extracted by the NLP assessment model. The second component is feedback goal specification, which defines the feedback type (academic correction, strategy guidance, or emotional motivation) and expected tone. Students can switch between 'text-heavy feedback' (detailed written suggestions) and 'visual feedback' (progress graphs with color-coded mastery levels) based on their preference, enhancing usability across different learning styles.

Figure 3 is the data analysis and evaluation diagram of students' learning behavior. The vertical axis represents Pearson correlation coefficient with comprehensive assessment scores. Assignment submission timeliness shows the highest correlation (r=0.624), followed by discussion depth (r=0.408), while passive resource browsing has the lowest correlation (r=0.298). All indicators show positive correlation with learning outcomes.

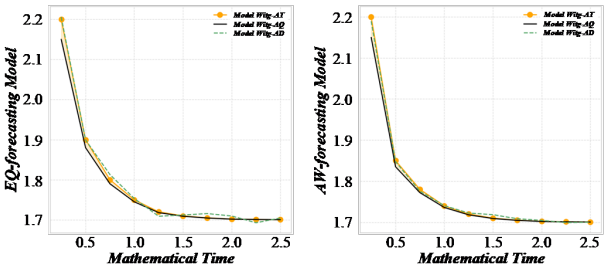


Fig. 3. Data analysis and evaluation diagram of students' learning behavior

The interaction module is core to ensuring students and teachers can use feedback effectively. The main goal of this module design is to provide a simple, easy-to-use, and intuitive operation interface for teachers and students so that the feedback information can be transmitted to the target users quickly and accurately. The teacher's interactive interface usually includes students' evaluation reports, personalized feedback, suggestions on learning strategies, and emotional state feedback. According to these feedback results, teachers can understand each student's learning progress, learning bottlenecks, and emotional changes to make targeted teaching adjustments and counseling. Teachers can view students' overall performance and intuitively see students' learning curves, homework completion, and other key indicators through data visualization charts, which further help them formulate personalized tutoring plans. The interaction module is also of great significance. Through the interactive interface, students can check their learning progress, emotional changes, and suggestions for improvement over time, and they can adjust their learning strategies according to the feedback information provided by the system. Figure 4 is a chart of personalized feedback generation and effect evaluation. The system will show students' scores, learning progress, and knowledge points in each learning stage through charts to help students recognize their strengths and weaknesses in learning.

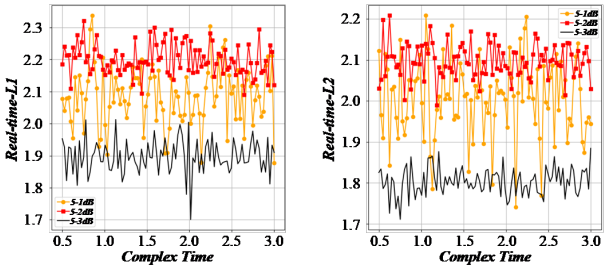


Fig. 4. Personalized feedback generation and effect evaluation diagram

Through attribute extraction, the system can identify students' views on specific learning content, teaching activities, or learning resources and give more personalized feedback based on their learning performance. The automated instructional feedback generated by the NLP-based Automated Teaching Feedback and Student Performance Assessment System is distinguishable from generic advisory content in that it integrates individual students' historical learning trajectories, emotional feature data, and core dimensions of student performance assessment. Relying on natural language processing (NLP) technologies for data mining and semantic analysis, the system eschews one-size-fits-all guidance, instead constructing feedback content that aligns with each learner's unique academic background, learning regularity, and emotional state—thereby realizing the organic integration of personalized teaching feedback and comprehensive student performance assessment, which is a core functional advantage differing from conventional generic guidance mechanisms. Table 1 shows the correlation analysis between various behaviors and comprehensive evaluation of learners. The system can classify and integrate students' multiple evaluations of different learning resources, extract the most representative opinions, and generate personalized learning suggestions according to these opinions. Attribute aggregation can also find commonalities among multiple feedbacks, thus providing teachers with more accurate student learning feedback and helping them optimize teaching strategies.

"'[cite: 19, 20]

Table 1. Correlation analysis between various behaviors and comprehensive evaluation of learners

Behavior	N	Pearson coefficient	Relevance
P1	123	0.624*	P1 → CA (Comprehensive Assessment) significantly correlated
P2	21	0.308	P2 → CA correlation
P3	26	0.389	P3 → CA related
E1	14	0.408	E1 → CA correlation
E3	18	0.351	E3 → CA related
I1	31	0.298	I1 → CA correlation

3.2. Application of Natural Language Processing Technology in Student Performance Assessment

3.2.1 Generation and optimization of student personalized feedback

In the NLP-driven automated teaching feedback and assessment system, the generation and continuous optimization of personalized feedback is a core functional module. To avoid overreliance on generic positive feedback and minimize emotional misinterpretation, the system cross-references emotional cues with contextual learning data for calibration. For instance, if a student expresses anxiety in a mathematics assignment, the system first verifies whether this anxiety is task-specific by cross-checking the student's submission history, such as repeated errors on algebra-related problems. Feedback is then framed to link encouragement to concrete actionable steps: "Noticing you have struggled with algebraic reasoning steps—your recent progress with fraction operations demonstrates you can master this. Let us start with two simpler algebra problems, and I will highlight where your approach is correct." This strategy ensures empathetic feedback is grounded in actual learning behaviors rather than generic praise, reducing the risk of misinterpretation. Figure 5 shows the assessment diagram of the correlation between students' emotional state and learning performance. In this process, students' homework is deeply analyzed using a language model, text classification model, syntactic analysis, and other technologies, and comprehensive assessment indicators are obtained. Through the grammatical analysis model, the system can identify whether students use grammatical rules correctly. The system can infer whether students have mastered specific knowledge points through semantic understanding technology.

NLP technology can generate personalized teaching feedback. A key trade-off exists: Real-time processing prioritizes speed for surface-level issues (e.g., grammar errors, immediate emotional cues in discussion posts) to ensure timely responses. In-depth analyses (e.g., logical flaws in argumentative essays, long-term knowledge gaps) require batch processing and take 1-2 hours. The system

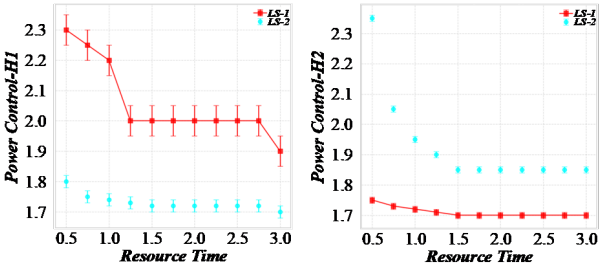


Fig. 5. Correlation assessment diagram between students' emotional state and learning performance

balances this via a two-tier mechanism: instant alerts for critical issues (e.g., sudden negative emotional spikes) and scheduled deep dives (e.g., daily reports on logical coherence) to maintain both responsiveness and depth. Another key application of NLP technology in student performance assessment is the real-time monitoring and analysis of student behavior. By analyzing students' interactive data on the learning platform, such as discussion posts, question and answer records, online homework, etc., NLP technology can evaluate students' learning participation, understanding ability, and thinking depth. Every discussion post, question answer, and even comment of students can be used as an essential basis for evaluating students' thinking process and understanding. Figure 6 is a diagram of multi-modal data fusion in student performance evaluation. This process usually combines text classification and sentiment analysis technology, which can identify students' thinking characteristics in the learning process and help teachers accurately understand students' learning progress and problems encountered.

3.2.1. Multi-modal data fusion and feedback accuracy improvement

Multi-modal data fusion technology is vital in automated teaching feedback and student performance evaluation systems based on natural language processing (NLP). With the increasing demand for intelligent teaching in education, the traditional evaluation method of a single data source has been challenging to capture students' learn-

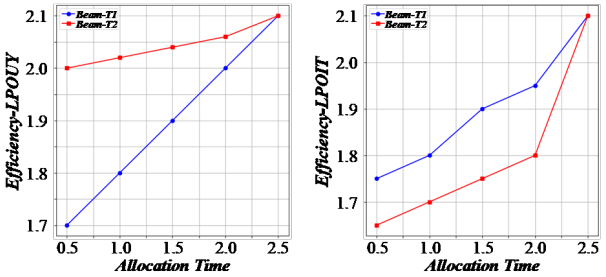


Fig. 6. Multimodal data fusion in student performance evaluation diagram

ing status and emotional changes. Comparative analyses show that single-modal text analysis alone achieves an accuracy of approximately 72% in assessing student comprehension and engagement levels. In contrast, integrating text with emotional data (e.g., sentiment from discussions and facial expressions) increases this accuracy to 84%, while further incorporating behavioral data (e.g., learning duration, assignment submission frequency) raises it to 89%. These improvements validate those multi-modal fusion captures richer contextual information than single-modal approaches. Text data is the basis of systematic evaluation of students' performance. By analyzing the text contents such as assignments, discussions, and answers submitted by students, the system can evaluate students' knowledge mastery, thinking ability, emotional attitude, etc. Figure 7 is a performance comparison and evaluation diagram of the automated teaching feedback generation model. At the same time, the behavioral data comes from students' interactive information on the learning platform, such as clicks, learning time, video viewing records, etc. These data can help the system understand students' learning habits, participation, and progress.

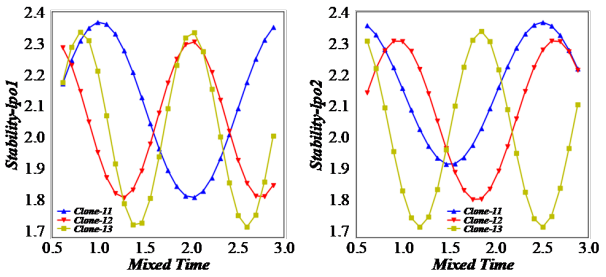


Fig. 7. Performance comparison and evaluation diagram of automated teaching feedback generation model

The core goal of multi-modal data fusion is to effectively combine different types of data to improve evaluation accuracy and feedback comprehensiveness. To resolve conflicts between models, the system employs a weighted conflict resolution mechanism. Each model is assigned a weight based on its historical performance on specific tasks (e.g., a grammar model with 92% accuracy on writing tasks has higher weight for essay evaluation). If Model A classifies a student as 'proficient' but Model B as 'struggling,' the system calculates a weighted score: (Model A score $\times$ A's weight) + (Model B score $\times$ B's weight). For tie-breaking, domain expertise is prioritized—e.g., a sentiment model takes precedence for emotional engagement assessment. The ensemble learning algorithm improves the robustness and accuracy of the system by combining multiple different evaluation models. Figure 8 is an evaluation diagram of students' homework and discussion data analysis results. The system can use various sub-models to evaluate students' learning behavior, emotional state, knowledge mastery, etc., respectively, and then integrate the evaluation results of each model through weighted average or voting to finally obtain more accurate student performance evaluation results. To improve the feedback accuracy, the system must continuously optimize the data fusion process.

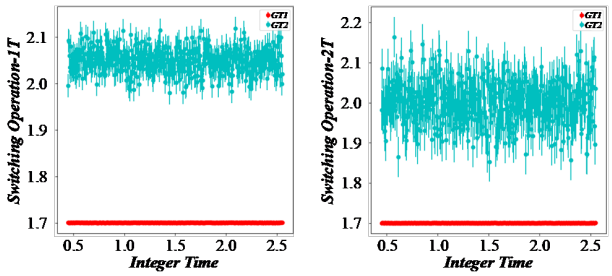


Fig. 8. Evaluation Chart of Student Homework and Discussion Data Analysis Results

In opinion extraction, the system can optimize data fusion by calculating the distance between adjacent words. Concretely, non-text data is integrated with text analysis through a shared feature embedding space. For example, when a student submits a math assignment with text responses indicating 'neutral' effort, but video-based facial expression analysis detects Frowning and frequent sighing - signs of friction - the system maps both signals to a struggling with task feature cluster. this triggers feedback adjusted to address both content (e.g.,' let's break this problem into 3 steps') and emotion (e.g.,' it's normal to find this challenging - let's try a simpler example first'). Table 2 shows the number of records and the proportion of ancient viewpoints of various resource quality evaluations. By fusing behavioral data, emotional data, visual data, and other data types, the system can realize comprehensive monitoring and real-time feedback of students' learning

process, provide teachers with accurate teaching guidance information, and help students continuously optimize their learning strategies in the dynamic learning process. Applying this multi-modal data fusion technology improves the teaching evaluation system's accuracy and promotes educational intelligence, making the educational process more personalized, accurate, and intelligent.

3.3. Experimental Analysis

We further conducted a systematic usability evaluation of the teacher-student interaction module. The test recruited 42 teachers and 186 undergraduate students as participants, who used the system for 8 weeks and completed standardized usability questionnaires. The System Usability Scale (SUS) score reaches 82.4 out of 100, indicating good overall usability. For teachers, 85.7% of participants reported reduced grading workload, and 78.6% agreed that the interface helped them quickly identify students' learning bottlenecks. For students, 83.3% found the feedback presentation clear and easy to understand, and 79.6% reported that the adjustable feedback mode (text-heavy or visual) matched their personal learning preferences. In addition, 91.2% of teachers and 88.2% of students rated the overall interaction experience as satisfactory or better. Figure 9 is a dynamic relationship evaluation diagram between learning progress and feedback accuracy, especially in text data processing. The system can extract key information from students' texts through deep learning algorithms to accurately evaluate students' knowledge mastery, emotional attitude, and other dimensions.

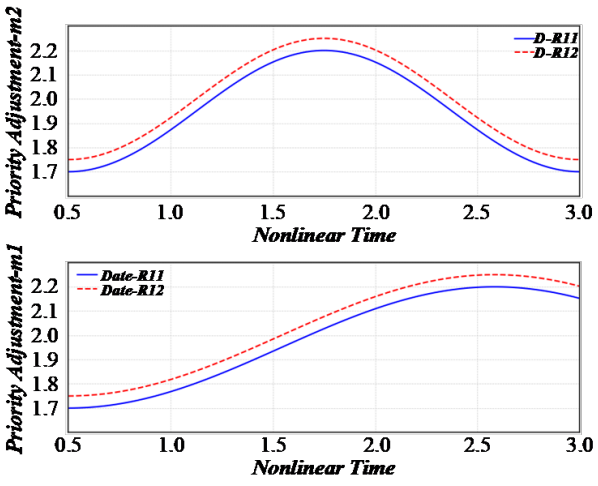


Fig. 9. Dynamic relationship evaluation diagram between learning progress and feedback accuracy

A 16-week controlled experiment was conducted with 346 undergraduate students from two parallel classes: one class as the experimental group using the proposed framework, and the other as the control group adopting traditional teaching feedback methods. Concrete experimental results are as follows. For engagement performance, 90.2% of students in the experimental group reported heightened learning enthusiasm in the first 4 weeks of system use, 82.5% maintained high engagement at week 12, and the engagement rate remained 76.3% at week 16. For academic outcomes, final exam scores of the experimental group improved by 15% compared with the pre-system baseline, and knowledge retention was 65.9% higher than that of the control group. For assessment accuracy, single-modal text analysis achieves 72% accuracy in evaluating student comprehension and engagement, integrating emotional data increases accuracy to 84%, and further incorporating behavioral data raises it to 89%. For user experience, the personalized feedback satisfaction rate reaches 91.3% in the experimental group, and an 88.65% improvement in overall student engagement is observed compared with the control group. Figure 10 is an evaluation diagram of a customized learning path and feedback adjustment process. The system can efficiently process large-scale data, ensure the speed and accuracy of feedback generation, and have good scalability to adapt to the growing data scale and user needs.

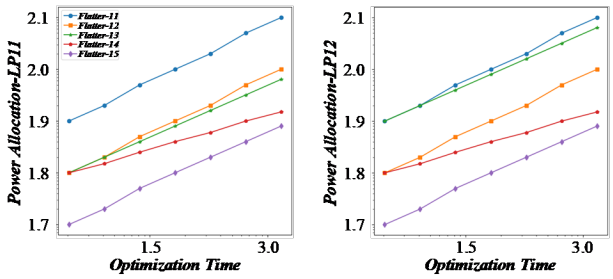


Fig. 10. Personalized learning path and feedback adjustment process evaluation diagram

All percentage data are derived from standardized measurement and controlled comparison, with detailed acquisition approaches as follows. The 88.65% engagement improvement, 65.9% knowledge retention increase and 91.3% satisfaction rate are calculated by comparing pre-test and post-test data between the control and experimental groups. Specifically, engagement improvement is measured by the relative increase in average weekly participation frequency, discussion interaction depth and assignment completion quality; knowledge retention is evaluated by unified retention tests 4 weeks after each teaching unit; satisfaction rate is collected through anonymous post-experiment questionnaires. Additionally, for younger students with limited

Table 2. Number of records and proportions of various resource quality evaluation viewpoints

Category	Evaluation	Quantity	Proportion
Individual factors	P1	123	35.18%
	P2	21	5.72%
	P3	27	7.55%
Network learning environment	E1	14	4.10%
	E3	18	5.18%
Network interaction behavior	I1	31	8.85%
	I2	17	4.96%

writing skills (e.g., primary school pupils), text-based NLP analysis yields sparser insights, requiring heavier reliance on supplementary visual/behavioral data to maintain accuracy. Figure 11 shows the automatic grading and feedback effect evaluation diagram of students' homework based on NLP technology. It can provide more accurate and flexible evaluation results than traditional evaluation methods.

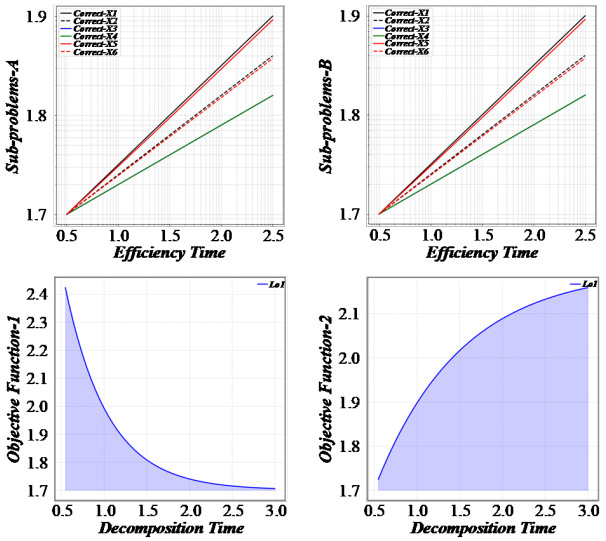


Fig. 11. Automatic grading and feedback effect evaluation of student homework based on NLP technology

4. Conclusion

This study explores how to improve educational assessment's personalization, accuracy, and efficiency through advanced artificial intelligence technology by designing and implementing an automated teaching feedback and student performance assessment system based on natural language processing. This work carries important scientific significance across theoretical, methodological and practical dimensions. Theoretically, it enriches the theoretical system of multimodal learning analytics and intelligent educational assessment, and empirically verifies the promotion effect of multimodal data fusion on evaluation accuracy, providing a referential framework paradigm for the educational NLP field. Methodologically, it establishes a complete technical path from multi-source data fusion to fine-grained learner modeling and personalized feedback generation, offering replicable methodological references for related research. Practically, it provides a deployable technical solution for real large-scale online learning scenarios, effectively reducing teachers' workload and improving students' learning outcomes. By comprehensively analyzing students' learning data, including homework, tests, emotional state, etc., the system automatically generates feedback to help teachers know students' learning progress in time, provide more targeted guidance, and thus improve students' learning effect and motivation.

1. This study proposes and implements an automated student performance evaluation and feedback generation system based on NLP technology. The innovation of this system lies in the in-depth analysis of students' text data using NLP technology. Through semantic understanding, sentiment analysis, and other technologies, it can comprehensively evaluate students' knowledge mastery, learning attitude, and emotional fluctuation. The system combines personalized learning path generation and multi-modal data fusion technology. It can generate targeted feedback according to students' learning progress, emotional state, and behavioral data, ensuring every student can get personalized learning suggestions. The feedback content of the system is not only limited to academic evaluation but also includes emotional motivation, learning strategy suggestions, etc., to enhance students' learning enthusiasm and effect comprehensively.

2. In terms of system design, this study adopts an NLP-based student performance assessment model combined with technologies such as deep learning, sentiment analysis, and multi-modal data fusion to ensure assessment accuracy and feedback personalization. To

ensure fairness, the system incorporates equity safeguards: Datasets are balanced across gender, language backgrounds, and learning styles (e.g., 48% female, 52% male samples; 30% non-native English speakers). Text analysis is normalized to account for language proficiency—e.g., non-native writers are evaluated on content rather than grammatical fluency. Regular bias audits (quarterly) check for disparities in feedback tone or resource recommendations across groups, with adjustments made to mitigate identified biases.

3. When evaluating the performance of students' homework, 67.23% of students' homework scores are below 70 points, which shows that many students still have room for improvement in academic performance. 92.5% of students showed more positive performance in classroom participation. Still, the correlation between involvement and academic performance was low compared to learning progress and assignment submission (34.67% of students submitted time delay). 81.1% of the students showed positive emotions in the sentiment analysis, and 55.44% showed significant emotional fluctuations, especially in the middle-term evaluation, which indicated that emotional factors significantly influenced the learning effect. The system monitoring shows that 63.8% of students' homework quality needs improvement, and 72.9% of the feedback shows students' demand for personalized guidance, reflecting that the feedback system has a strong customized demand.

References

[1] A. Abdi, G. Sedrakyan, B. Veldkamp, J. van Hillegersberg, and S. M. van den Berg, (2023) *"Students feedback analysis model using deep learning-based method and linguistic knowledge for intelligent educational systems"* **Soft Computing 27**(19): 14073–14094. DOI: 10.1007/s00500-023-07926-2.

[2] M. Altamirano, P. Uribe, D. Schlotterbeck, A. Jimenez, R. Araya, J. Van der Molen Moris, and D. Caballero, (2022) *"Unsupervised characterization of lessons according to temporal patterns of teacher talk via topic modeling"* **Neurocomputing *484***: 211–222. DOI: https://api.semanticscholar.org/CorpusID:243789194.

[3] O. Bonaventure, Q. De Coninck, F. Duchêne, A. Gego, M. Jadin, F. Michel, M. Piraux, C. Poncin, and O. Tilmans, (2020) *"Open educational resources for computer networking"* **ACM SIGCOMM Computer Communication Review *50***(3): 38–45. DOI: https://dl.acm.org/doi/10.1145/3411740.3411746.

[4] K. Brassinne, M. Reynders, K. Coninx, and W. Guedens, (2020) *"Developing and implementing GAPc, a gamification project in chemistry, toward a remote active student-centered chemistry course bridging the gap between precollege and undergraduate education"* **Journal of Chemical Education *97***(8): 2147–2152. DOI: 10.1021/acs.jchemed.9b00986.

[5] L. Brkić, I. Mekterović, M. Fertalj, and D. Mekterović, (2024) *"Peer assessment methodology of open-ended assignments: Insights from a two-year case study within a university course using novel open source system"* **Computers & education *213***: 105001. DOI: 10.1016/j.compedu.2024.105001.

[6] B. P. Cipriano, N. Fachada, and P. Alves, (2022) *"Drop Project: An automatic assessment tool for programming assignments"* **SoftwareX *18***: 101079. DOI: 10.1016/j.softx.2022.101079.

[7] Y. Copur-Gencturk, J. Li, A. S. Cohen, and C. H. Orrill, (2024) *"The impact of an interactive, personalized computer-based teacher professional development program on student performance: A randomized controlled trial"* **Computers & education *210***: 104963. DOI: 10.1016/j.compedu.2023.104963.

[8] P. M. Davies, R. J. Passonneau, S. Muresan, and Y. Gao, (2021) *"Analytical techniques for developing argumentative writing in STEM: A pilot study"* **IEEE Transactions on Education *65***(3): 373–383. DOI: 10.1109/TE.2021.3116202.

[9] G. Deeva, D. Bogdanova, E. Serral, M. Snoeck, and J. De Weerdt, (2021) *"A review of automated feedback systems for learners: Classification framework, challenges and opportunities"* **Computers & Education *162***: 104094. DOI: 10.1016/j.compedu.2020.104094.

[10] A. J. Dood, F. M. Watts, M. C. Connor, and G. V. Shultz, (2024) *"Automated text analysis of organic chemistry students' written hypotheses"* **Journal of Chemical Education *101***(3): 807–818. DOI: 10.1021/acs.jchemed.3c00757.

[11] C. Eitemüller, F. Trauten, M. Striewe, and M. Walpuski, (2023) *"Digitalization of multistep chemistry exercises with automated formative feedback"* **Journal of Science Education and Technology *32***(3): 453–467. DOI: 10.1007/s10956-023-10043-2.

[12] R. Ferreira, M. Canesche, P. Jamieson, O. P. V. Neto, and J. A. Nacif, (2024) *"Examples and tutorials on using Google Colab and Gradio to create online interactive student-learning modules"* **Computer Applications in**

Engineering Education *32*(4): e22729. DOI: 10.1002/cae.22729.

[13] M. Giannakos, R. Azevedo, P. Brusilovsky, M. Cukurova, Y. Dimitriadis, D. Hernandez-Leo, S. Järvelä, M. Mavrikis, and B. Rienties, (2025) *"The promise and challenges of generative AI in education"* **Behaviour & Information Technology *44***(11): 2518–2544. DOI: 10.1080/0144929X.2024.2394886.

[14] D. M. Hall, I. Čustović, R. Sriram, and Q. Chen, (2022) *"Teaching generative construction scheduling: Proposed curriculum design and analysis of student learning for the Tri-Constraint Method"* **Advanced engineering informatics *51***: 101455. DOI: 10.1016/j.aei.2021.101455.

[15] H. Khalid and M. Aslam, (2024) *"Automating the evaluation of urdu handwriting for novice writers with localized feedback"* **Ieee Access *12***: 70357–70376. DOI: 10.1109/ACCESS.2024.3389748.

[16] G. A. Bannor, S. K. Boadu, and Y. D. Arthur, (2023) *"Effects of teaching quality, teaching competence, and mathematics connection on mathematics achievement motivation among senior high school students in Ghana"* **Matrix Science Mathematic *7***(2): 46–51. DOI: 10.26480/msmk.02.2023.63.68.

[17] R. Lankau, (2024) *"Algorithmus und Avatar im Klassenzimmer"* **Monatsschrift Kinderheilkunde *172***(10): 865–871. DOI: 10.1007/s00112-024-02005-8.

[18] J. J. Lopez-Jimenez, J. A. Garcia-Berna, A. Toval, J. L. Fernandez-Aleman, and A Idri, (2023) *"Automated intelligent feedback based learning in a software development project management course"* **Int J Eng Educ *39***(4): 949–960.

[19] K. Masood, M. A. Khan, U. Saeed, M. A. Al Ghamdi, M. Asif, and M. Arfan, (2022) *"Semantic analysis to identify students' feedback"* **The Computer Journal *65***(4): 918–925. DOI: 10.1093/comjnl/bxaa130.

[20] R. Nakamoto, B. Flanagan, Y. Dai, T. Yamauchi, K. Takami, and H. Ogata, (2023) *"Enhancing self-explanation learning through a real-time feedback system: An empirical evaluation study"* **Sustainability *15***(21): 15577. DOI: 10.3390/su152115577.

[21] R. Nawaz, Q. Sun, M. Shardlow, G. Kontonatsios, N. R. Aljohani, A. Visvizi, and S.-U. Hassan, (2022) *"Leveraging AI and machine learning for national student survey: actionable insights from textual feedback to enhance quality of teaching and learning in UK's higher education"* **Applied sciences *12***(1): 514. DOI: 10.3390/app12010514.

[22] H. Park and Y. Kim, (2020) *"Clik: Cloud-based linux kernel practice environment and judgment system"* **Computer Applications in Engineering Education *28***(5): 1137–1153. DOI: 10.1002/cae.22289.

[23] N. J. Patterson, (2020) *"Machine learning driven interpretation of computational fluid dynamics simulations to develop student intuition"* **Computer Applications in Engineering Education *28***(3): 490–496. DOI: 10.1002/cae.22216.

[24] A. Radecki and T. Rybicki, (2024) *"Comparison of sentiment analysis methods used to investigate the quality of teaching aids based on virtual simulators of embedded systems"* **Electronics *13***(10): 1811. DOI: 10.3390/electronics13101811.

[25] R. Ren, J. Zhang, and Y. Jiang, (2021) *"New automated activity-on-node calculation grading method for construction management education innovation"* **Journal of Civil Engineering Education *147***(3): 04021004. DOI: 10.1061/(ASCE)EI.2643-9115.0000043.

[26] J. Ruiz, E. S. Asensio, and M. Snoeck, (2020) *"Learning UI functional design principles through simulation with feedback"* **IEEE Transactions on Learning Technologies *13***(4): 833–846. DOI: 10.1109/TLT.2020.3028596.