Advances in Medical Education and Practice

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Development and External Validation of a Machine Learning Model for Automated Feedback Quality Assessment in Chinese Anesthesiology Residency Training

Lifeng Yao, Yijun Chen, Jing Shen, Junge Zhang, Yiwei Zhang, Guojin Liang, Yiqin Ji

Department of Anesthesiology, The First Affiliated Hospital of Ningbo University, Ningbo, Zhejiang, People's Republic of China

Correspondence: Yijun Chen, Department of Anesthesiology, The First Affiliated Hospital of Ningbo University, No. 59 Liuting Street, Haishu District, Ningbo, Zhejiang, 315010, People's Republic of China, Email fyychenyijun@nbu.edu.cn

Purpose: High-quality narrative feedback is essential for competency-based medical education, but manual evaluation of feedback is time-consuming and subjective. This research aims to develop and validate a machine learning (ML)-based model to automate the bulk evaluation of feedback quality from anesthesiology residency program instructors.

Methods: Using 990 narrative feedback entries from October 2023 to November 2025 at the First Affiliated Hospital of Ningbo University, we conducted training and validation. An additional 587 feedback records from Ningbo Li HuiLi Hospital were used as an external test set. Text processing employed the jieba Chinese word segmenter combined with an anesthesia-specific vocabulary database to extract TF-IDF and manual features. Data imbalance was addressed using the Synthetic Minority Oversampling Technique (SMOTE). Logistic regression (LR), random forests (RF), and Gradient Boosting Machine (GBM) were used for training and validation. Model performance was measured using the area under the receiver operating characteristic curve (AUC-ROC), accuracy, cross-validation accuracy, precision, recall, and F1 score.

Results: In internal training, LR performed optimally, demonstrating the best overall performance (F1 score: 0.941) and stability (cross-validation accuracy: 0.925 ± 0.026), along with the highest precision (0.906). In external testing, the LR model achieved an overall accuracy of 0.840 (95% CI: 0.808–0.867), with high recall (0.956) and moderate precision (0.636) for identifying high-quality feedback, yielding an F1 score of 0.764 and an AUC of 0.729.

Conclusion: This study successfully developed and externally validated a machine learning-based model for automated feedback quality assessment in Chinese anesthesiology residency training. With its high recall and stable internal performance, the model may serve as a screening tool to support competency-based medical education by enabling batch evaluation of narrative feedback.

Keywords: machine learning, natural language processing, medical education, educational improvement, feedback quality

Introduction

The competency-based residency training assessment system generates a large volume of textual data in the form of narrative feedback.[1] In anesthesiology residency training, the application of knowledge, clinical skills, and rapid decision-making abilities is critical. High-quality, timely, and specific feedback plays a vital role in correcting errors, reinforcing proper behaviors, and improving situational judgment and clinical judgment skills. Criteria for high-quality feedback can be summarized as six key characteristics: centered on resident physician conduct, providing detailed information, including negative feedback, evaluating professional competence and communication skills, being targeted to specific actions, and identifying areas for improvement.[2,3] The shift from time-based training to competency-based medical education (CBME), exemplified by frameworks such as the Accreditation Council for Graduate Medical Education (ACGME) Milestones and Entrustable Professional Activities (EPAs), has placed new demands on feedback quality, which now serves as primary evidence for documenting trainee progress and ensuring patient safety.[4–6]

https://doi.org/10.2147/AMEP.S599543
Received: 28 January 2026
Accepted: 23 April 2026
Published: 30 April 2026

However, traditional manual methods for reviewing feedback data are time-consuming, subjective, and prone to bias. It is challenging to manage the difficulties presented by large volumes and high data complexity, resulting in a significant amount of valuable feedback data going underutilized.[1,7] With the growing popularity of mobile evaluation apps, the volume of feedback requiring assessment has surged exponentially, making these issues even more pronounced. Researchers have employed machine learning tools, such as Natural Language Processing (NLP), to improve the efficiency of reviewing textual feedback data from pre-clinical medical students.[8] Additionally, research has shown the efficacy of machine learning in feedback analysis for certain general education scenarios.[9–11] However, research explicitly targeting feedback quality in anesthesiology residency training remains extremely limited. This gap is particularly pronounced in the Chinese medical context, where specialized text segmentation is required due to ambiguous word boundaries and dense terminology—factors that general NLP tools often fail to handle accurately.[12,13]

To address these challenges, this study aims to develop and validate a machine learning-based model. Through incorporating a specialized Chinese text tokenizer combined with an anesthesia terminology database, the model's processing functions for Chinese text are enhanced. The research verifies whether this model can automatically predict the quality of feedback provided by anesthesia residency training supervisors in batches. To our knowledge, this represents the first attempt to develop and validate a dedicated ML assessment tool for narrative feedback within the context of Chinese anesthesiology residency education.

Materials and Methods

Sample Size Estimation

Based on the Sample Size Calculation Method for Predictive Models proposed by Riley et al,[14] for 30 candidate prediction parameters (including 12 TF-IDF features, six manual features, and 12 keyword features), assuming an expected high-quality feedback rate of 25%, and adopting a conservative model performance expectation ($R^2 = 0.15$), the calculated total sample size required is approximately 907.

General Information

Using narrative feedback data from October 2023 to November 2025, provided by faculty instructors in the Department of Anesthesiology at Ningbo University Affiliated First Hospital, extracted from the hospital's teaching management system. This includes free-text feedback from three assessment tools: the Resident Monthly Performance Evaluation Forms, the Mini Clinical Evaluation Exercise (Mini-CEX), and the Direct Observation of Procedural Skills (DOPS). After manual review to exclude non-anesthesiology faculty feedback, the final dataset comprised 990 valid feedback entries. External testing selected the Department of Anesthesiology at Ningbo Li HuiLi Hospital. Data extraction sourced 587 free-text entries from the China Medical Continuing Education Television (CCMTV) system, specifically from 360 evaluations, routine assessments, and departmental exit examinations.

Ethics

This study received approval from the ethics committee, which granted exemption from review and waived informed consent requirements (Approval No.: Ningbo University Affiliated First Hospital Ethics Review 2025 Research No. 321A-01).

Establishing Evaluation Criteria

Based on Teaching Literature,[3] Quality scores were assigned according to the criteria (Table 1), with high-quality feedback features deemed present or absent. Among the six features, ≥ four features were considered high-quality. Taking "Clear process for fiberoptic bronchoscope-guided double-lumen endotracheal intubation, thorough preoperative assessment, strict aseptic technique, smooth procedure, and post-procedure assessment was accurate". This example demonstrates the presence of four characteristics beyond "negative feedback and identification of areas for improvement," thus qualifying as high-quality feedback.

Table 1 Six Key Characteristics of High-Quality Feedback: Definition and Examples

High-Quality Feedback Features	Definition	Example
1. Centered on the conduct of resident physicians	Evaluating a specific behavior (not personality or traits) of a resident physician.	Mastery of endotracheal intubation skills.
2. Provide detailed information	Provide information describing observed circumstances or actions taken (regardless of length).	Accurate determination of catheter position through auscultation of breath sounds.
3. Negative feedback	Point out the shortcomings or mistakes of resident physicians.	Weak foundation in anesthesiology.
4. Evaluate professional competence and/or communication skills	Assess whether resident physicians possess the planning, preparation, and communication skills required for professional practice.	The resident demonstrated excellent overall performance, strictly adhering to indications, thoroughly completing preoperative preparations, and appropriately managing postoperative care.
5. Targeted	Directly point out specific information related to the specific actions of the resident physician.	Comprehensive airway assessment during preoperative visit.
6. Identify areas for improvement	Identify specific areas where resident physicians need to strive for improvement.	Humanistic care can be strengthened.

Methods

The study was conducted according to the Declaration of Helsinki principles and reported following the SQUIRE - EDU guidelines (see Supplementary Table 1). The technical workflow comprised the following key steps, as illustrated in Figure 1. (1) Text Preprocessing: Anonymize information regarding instructors, trainees, and patients by removing identifiable details (dates, uncommon procedures, etc), irrelevant content (special characters, redundant spaces, etc), and common Chinese stop words (eg., functional characters), and spelling corrections based on clinical terminology standards. (2) Manual Labelling: Each feedback label is manually rated by two independent, experienced experts (A and B) according to Table 1 to indicate feedback quality. Differences are adjudicated by a third expert (C, a deputy chief physician and higher with ≥10 years of teaching experience), and the results are compiled into an Excel file. (3) Feature Engineering: Maximum feature count was set to 500 with n-gram range 1–2 to prevent overfitting while maintaining interpretability. We extracted TF-IDF text features and manual features (text length, word count, domain-specific terminology count, average word length), then merged them into a comprehensive feature matrix. Regularization was subsequently applied through the logistic regression model's built-in L2 penalty. (4) SMOTE was applied with k_neighbors = 5 to generate synthetic minority samples. (5) ML Model Construction (Figure 2): Three machine learning algorithms (LR, RF, GBM) were employed; The dataset was randomly split into training (70%) and validation (30%) sets with stratification to preserve class distribution; Visual Studio Code (version: 1.106.3) was used to develop a GUI

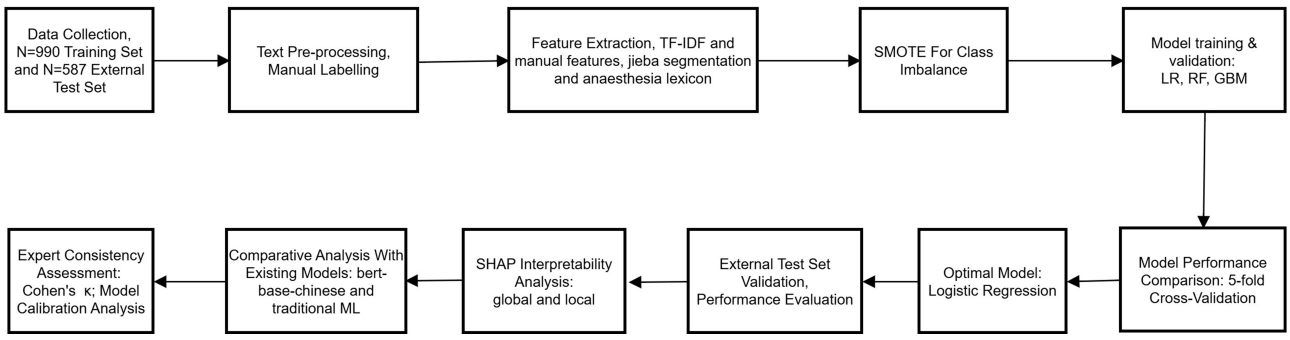

Figure 1 Overall Flowchart for Constructing and Validating the Quality Evaluation Model for Anesthesiology Residency Training Feedback.

https://doi.org/10.2147/AMEP.S599543

```
jieba.load_userdict('anaesthesia_lexicon.txt')
vectorizer = TfidfVectorizer(max_features=500, ngram_range=(1,2))
X_text = vectorizer.fit_transform(corpus)
X_hand = df[['length','prof_count','avg_len']].values
X = hstack([X_text, X_hand])
```

Figure 2 Key Python code segments for feature engineering, model training, and assessment criteria, implementing jieba Chinese word segmentation enhanced with an anaesthesia-specific lexicon.

analysis program supporting batch processing of Chinese text feedback (Figure 3). The "jieba Chinese word segmenter" was employed, supplemented with the "Medical Anesthesiology Vocabulary Database." (6) Model assessment criteria: Model effectiveness is assessed using the AUC-ROC, accuracy, cross-validation accuracy, precision, recall, and F1 score. Accuracy is the proportion of correctly predicted samples among all samples; values closer to 1 are better. Five-fold cross-validation with stratified sampling was used for model comparison; higher values indicate more stable classification performance and stronger generalizing power. Precision is the proportion of actual positive samples among all samples predicted as positive. Recall is the proportion of all actual positive samples correctly predicted as positive, with values closer to 1 being better. The F1 score is the harmonic mean of exactness and recall, with higher values indicating better model effectiveness. (7) External Testing: After the model makes predictions on the test set, analysis results are automatically exported and saved as structured Excel files for the following statistical analysis and visualization. (8)

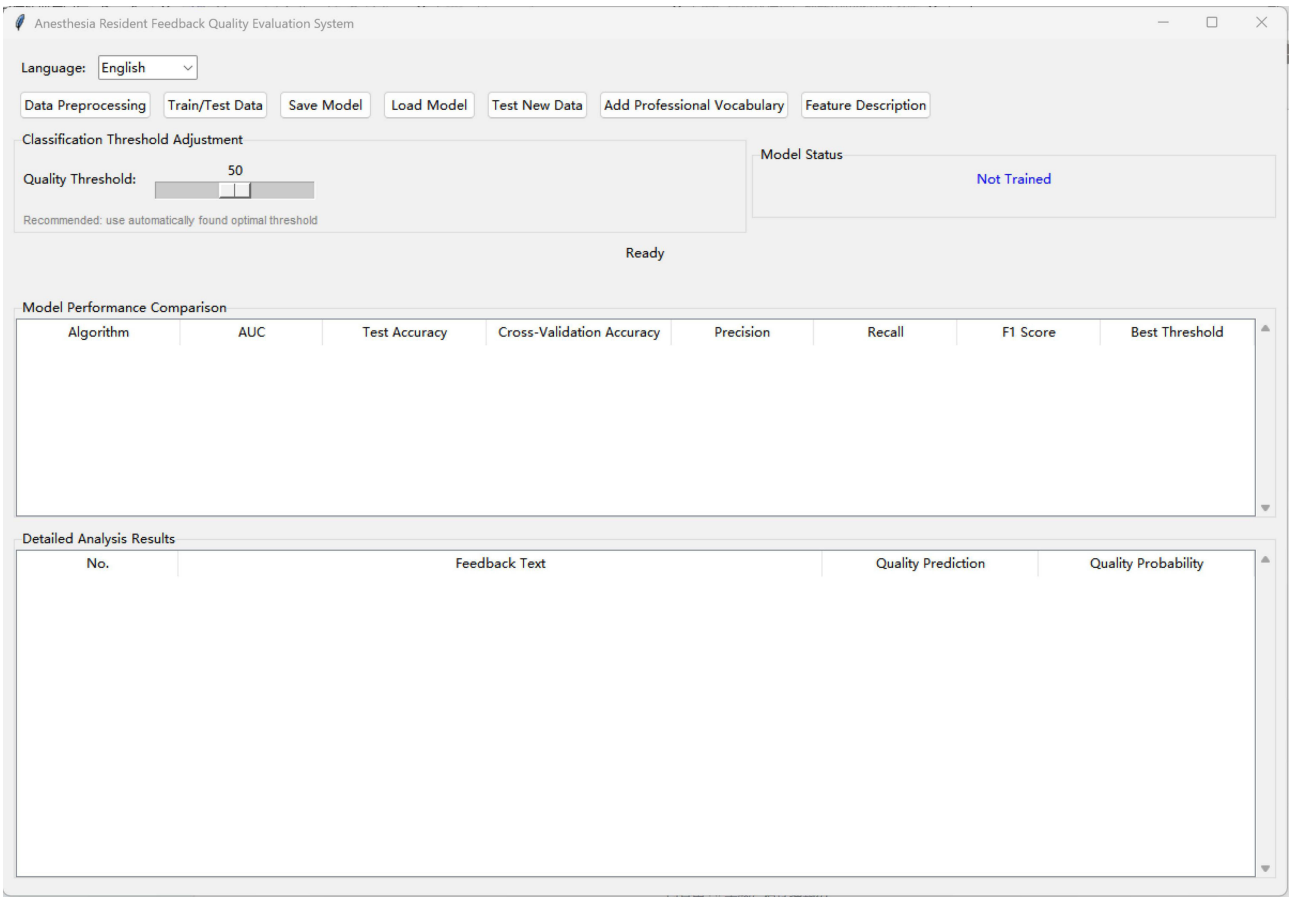

Figure 3 Main interface of the ML-based feedback quality analysis system for anesthesiology residency training. The GUI integrates data preprocessing, model training/testing, threshold adjustment, performance comparison, and individual feedback analysis modules, enabling non-technical users to evaluate feedback quality efficiently.

https://doi.org/10.2147/AMEP.S599543

The model's predictions are explained using SHAP analysis, its performance is benchmarked against prevalent models (BERT-base-Chinese and traditional ML), and its reliability is confirmed by expert consistency assessed by Cohen's kappa coefficient. (9) Model Calibration Analysis: Use calibration curves with quantile-based binning (10 bins, equal sample sizes) to assess probability reliability. Calculate the Brier score to quantify calibration performance. Compare predicted probabilities against observed positive fractions, with the diagonal representing perfect calibration.

Statistical Analysis

Continuous variables are expressed as mean ± standard deviation, while categorical variables are presented as frequency and percentage. Independent samples t-tests were used for quantitative data; appropriate statistical tests were used to compare model effectiveness metrics, with the significance level set at $P < 0.05$. All analyses were carried out using Python 3.8 (programming language environment) and the scikit-learn (ML), pandas (scientific computing), and numpy (data analysis) libraries.

Results

Dataset Feedback Feature Distribution

The training set (N = 990) had an average feedback length of 8.68 characters, significantly shorter than the test set (N = 587) at 12.95 characters ($P < 0.001$). The proportion of high-quality feedback was higher in the test set (40.20% vs. 24.44%), indicating overall superior quality of the test set compared to the training set. Professional vocabulary density was 0.22% in the training set hospitals, while none was detected in the test set, with a statistically significant difference ($P = 0.001$). The proportion of blank responses was similar across datasets (Training: 57.7%, Test: 52.6%, $P = 0.058$). To conclude, the test set exhibited characteristics of being "longer, better, and more colloquial," consistent with expectations for cross-institutional external validation. This also indicates the model keeps a strong generalization ability on more expressive texts.

Model Performance Comparison

To determine the optimal model, 5-fold cross-validation was applied to the training set to compare LR, RF, and GBM. As shown in Table 2, all three models showed excellent performance on the training set. LR achieved the best balance between overall performance (F1 score: 0.941) and stability (cross-validation accuracy: 0.925 ± 0.026), while also yielding the highest precision (0.906). Although RF and GBM achieved a recall of 1.000, their precision (both 0.875) was significantly lower than LR's, denoting a tendency to predict more samples as high quality and posing an overfitting risk. Considering the model's generalization capability, robustness, and the balance between precision and recall, LR was ultimately selected as the final model for external validation.

External Validation Results

The LR model was tested with 587 independent samples from another hospital (Table 3). The model attained an overall accuracy of 0.840 (95% CI: 0.808–0.867). While continuing with an extremely high recall rate (0.956), the precision rate was 0.636, yielding an F1 score of 0.764. The model reached 1.000 accuracy for identifying blank feedback. For the core task of quality grading non-blank feedback, accuracy was 0.662 (95% CI: 0.604–0.715) with an AUC of 0.729. The test

Table 2 Performance Comparison of Different ML Algorithms in Training Models

Algorithm	AUC	Accuracy Rate	Cross-validation Accuracy	Precision	Recall	F1 Score	Optimal Threshold
LR	0.953	0.929	0.925 ± 0.026	0.906	0.980	0.941	0.24
RF	0.955	0.917	0.915 ± 0.019	0.875	1.00	0.933	0.38
GBM	0.966	0.917	0.909 ± 0.019	0.875	1.00	0.933	0.24

Note: AUC refers to the area under the receiver operating characteristic curve (AUC-ROC).

Table 3 Performance of the LR Model on the External Test Set

	Project	Numerical Value	Note
Overall distribution of the test set	Total sample size	587	All feedback from the external test set
	High-Quality Feedback	236 (40.2%)	Feedback that meets high-quality standards
	Low-quality feedback	351 (59.8%)	Feedback that does not meet high-quality standards
	Blank feedback	309 (52.6%)	Feedback lacking substantive content
Confusion matrix	True positive (TP)	152	Actual high quality, predicted high quality
	False positive (FP)	87	Actual low quality, predicted high quality
	True negative (TN)	341	Actual low quality, predicted low quality
	False negative (FN)	7	Actual high quality, predicted low quality
Core performance metrics	Overall accuracy rate	0.840 (95% CI: 0.808, 0.867)	(TP + TN)/Total Sample Size
	Precision	0.636 (152/239)	TP/(TP+FP)
	Recall/Sensitivity	0.956 (152/159)	TP/(TP+FN)
	F1 Score	0.764	2 × Precision × Recall/(P+R)
Non-blank feedback performance only	Number of non-blank feedback samples	278	Valid feedback after removing blanks
	Non-blank feedback accuracy rate	0.662 (95% CI: 0.604, 0.715)	(TP + TN_non-blank)/Number of non-blank samples
	Non-blank feedback AUC	0.729	Model Discrimination Capability

set comprised 278 non-blank and 309 blank feedback entries. The confusion matrix shows: true positives (TP) = 152, false positives (FP) = 87, true negatives (TN) = 341, false negatives (FN) = 7.

SHAP Interpretability Analysis

The Shapley Additive exPlanations (SHAP) method was applied for post-hoc interpretation of the optimal LR model to explain its decision-making process in classifying "high-quality feedback." The analysis was conducted at both global and local levels.

Global Interpretation (Feature Importance and Direction)

The global importance of features was ranked based on their mean absolute SHAP values across all samples in the external test set (Table 4). The top three features—feature 20 (Specific Operational Description of Bronchoscopy), feature 11 (Level of detail/word density), and feature 17 (Preoperative assessment is thorough)—collectively accounted for 53.7% of the total predictive importance, indicating that the model's decisions are highly concentrated on a small set of semantically clear features.

Directional analysis of the mean SHAP values revealed how each feature typically influences the prediction: Positive contributors (Mean SHAP > 0): Features such as feature 17 (+0.117) and feature 1 (Text Length) (+0.075) consistently increased the predicted probability of feedback being classified as high-quality. This is consistent with educational theory, where in-depth review and substantive content are hallmarks of valuable feedback. Negative contributors (Mean SHAP < 0): Features like feature 11 (−0.083) and feature 13 (Negative Feedback_Exists) (−0.186) tended to decrease the quality score. Notably, the negative impact of feature 11 suggests that an excessively high density of certain vague terms may be perceived by the model as indicative of lower-quality, non-specific commentary.

Table 4 SHAP Global Feature Importance Ranking for the Logistic Regression Model (Top 15 Features)

Ranking	Feature ID	Characteristics	Average SHAP±SD	Average SHAP	Direction
1	feature 20	Specific Operational Description of Bronchoscopy	0.623 ± 0.731	−0.007	Negative
2	feature 11	Level of detail (word density)	0.535 ± 0.684	−0.083	Negative
3	feature 17	Preoperative assessment is thorough.	0.510 ± 0.676	+0.117	Positive
4	feature 26	Targeted_Existence	0.428 ± 0.587	−0.041	Negative
5	feature 6	Professional Vocabulary Count	0.354 ± 0.399	−0.004	Negative
6	feature 13	Negative Feedback_Exists	0.309 ± 0.319	−0.186	Negative
7	feature 12	Assessment in place_Existence	0.271 ± 0.323	−0.045	Negative
8	feature 1	Text Length (Characters)	0.266 ± 0.303	+0.075	Positive
9	feature 24	Operational Procedure Completeness	0.260 ± 0.297	+0.015	Positive
10	feature 16	Communication Skills_Existence	0.259 ± 0.309	+0.027	Positive
11	feature 28	Humanistic Care_Existence	0.197 ± 0.245	+0.043	Positive
12	feature 25	Sterile Technique_Existence	0.193 ± 0.231	−0.055	Negative
13	feature 7	Sterile Technique_Counting	0.193 ± 0.229	−0.076	Negative
14	feature 15	Planned Existence	0.184 ± 0.218	+0.030	Positive
15	feature 10	Postoperative Follow-up_Existing	0.183 ± 0.220	−0.011	Negative

Note: The direction column "Positive/Negative" indicates the average direction of contribution of this feature to the predicted value of "high-quality feedback.".

Local Interpretation (Explaining Individual Predictions)

Sample-Level Explanations (Waterfall Plots)

Figure 4A–C presents SHAP waterfall plots for three representative feedback entries. These plots visually decompose the model's prediction for a single sample, showing how each feature's SHAP value pushes the base prediction (E[f(X)] = 0.340) towards or away from the final outcome. This fine-grained view validates that the model's decisions for individual cases are driven by interpretable features.

Feature Behavior and Thresholds (Dependence Plots)

SHAP dependence plots (Figure 5A–C) illustrate how the model's response to a key feature changes with its value. For instance:

Feature 20 (Figure 5A): Its contribution enters a high-impact plateau once its TF-IDF value exceeds approximately 0.25, suggesting this value could serve as an objective threshold for identifying "specific procedural descriptions."

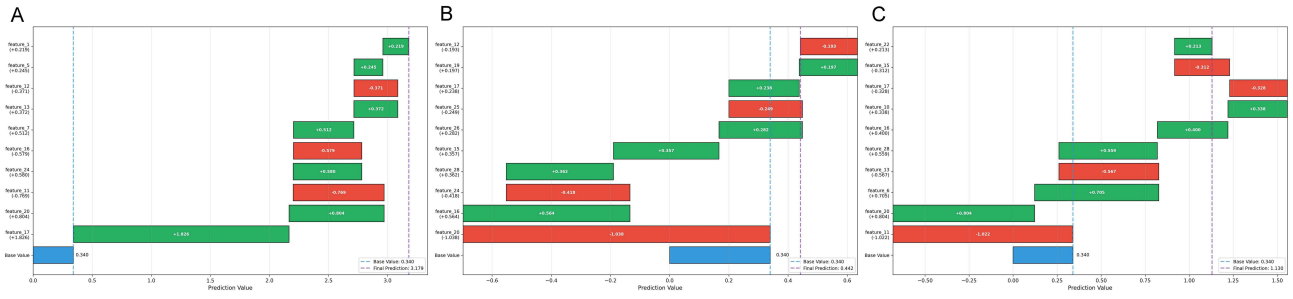

Figure 4 (**A–C**) SHAP waterfall plots demonstrating sample-level explainability.

https://doi.org/10.2147/AMEP.S599543

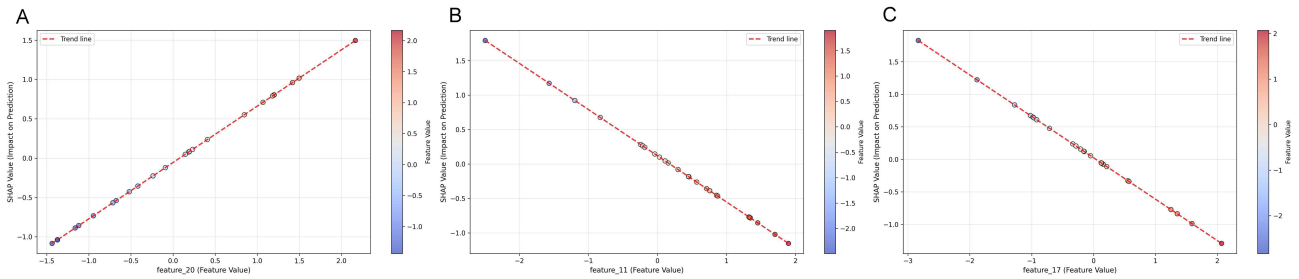

Figure 5 (**A**–**C**) SHAP dependence plots for top features.

Feature 11 (Figure 5B): Exhibits a near-linear negative relationship, where increasing term frequency steadily decreases the quality score.

Feature 17 (Figure 5C): Shows a slight inverted U-shaped relationship, implying that moderate keyword usage is optimal, while very high frequencies may not add incremental value.

Summary and Cumulative Impact

Cumulative Contribution

The cumulative contribution of the top 15 features reached 79.4% (Figure 6), demonstrating that the model relies on a compact, interpretable feature set rather than a diffuse array of weak signals.

Feature Summary Plot (Beeswarm Plot)

Figure 7 provides a holistic view of the top 15 features' effects across all samples. Each dot represents a sample, colored by the feature's value. This plot confirms the trends identified in Table 4 and Figure 4, such as the high value and positive impact of feature 17 (red/pink dots clustered in high SHAP region) and the broadly distributed negative impact of feature 11.

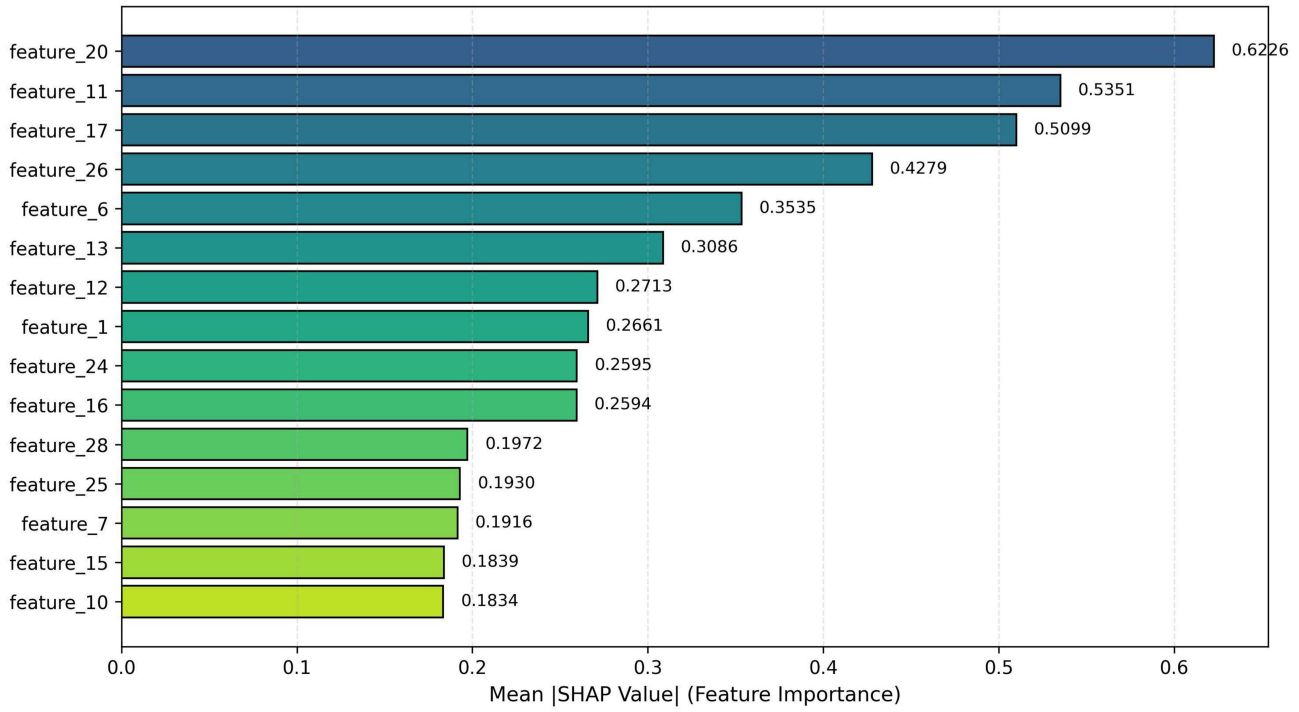

Figure 6 Cumulative SHAP value contribution of the top 15 features.

 https://doi.org/10.2147/AMEP.S599543

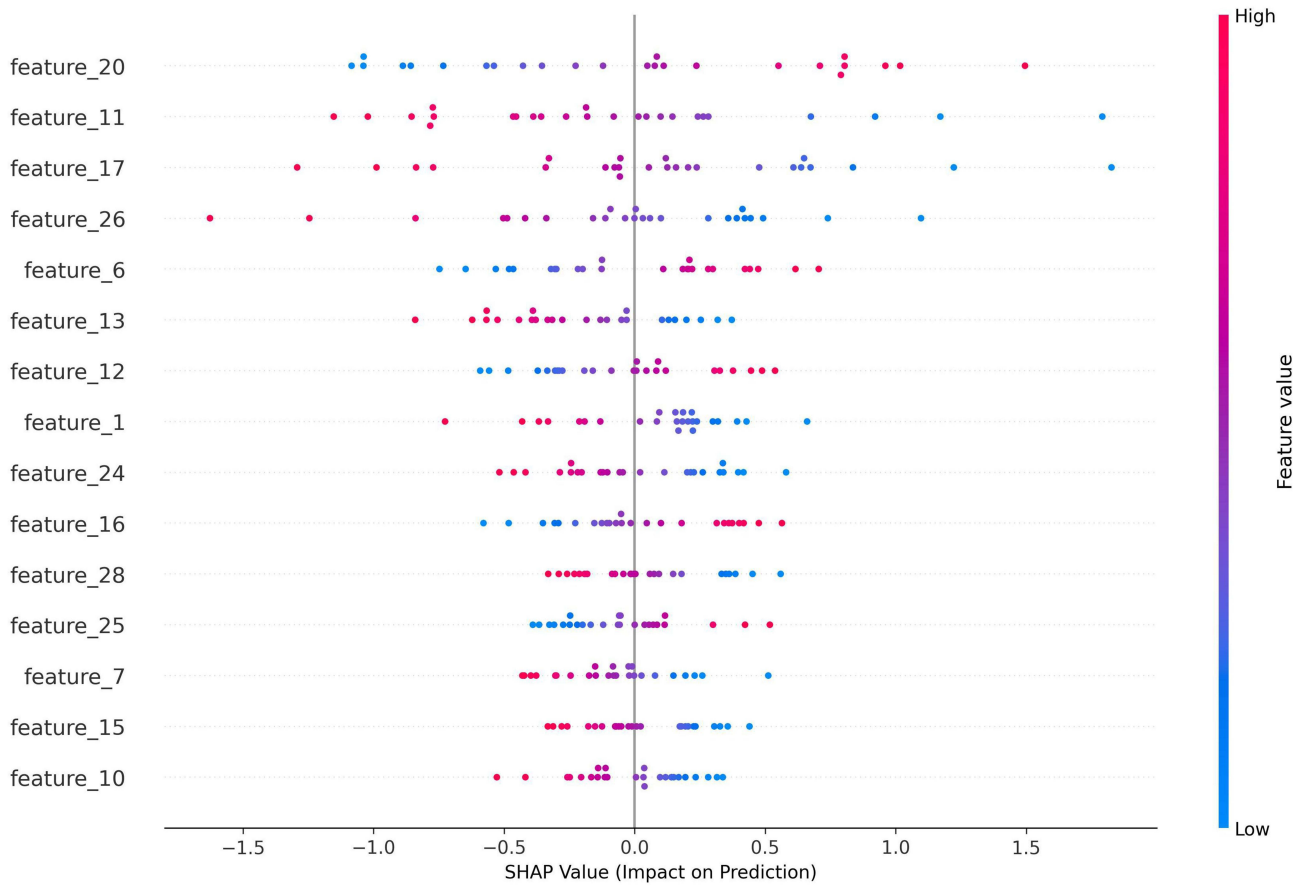

Figure 7 SHAP summary plot (beeswarm) for the top 15 features.

Educational Interpretability and Model Trust

The SHAP analysis effectively connects the gap between the model's statistical output and actionable educational insights. By identifying that features like "thorough preoperative assessment" and "specific procedural description" are primary positive drivers, the model provides instructors with evidence-based, concrete levers for improving their feedback quality. Conversely, flagging "vague term density" as a key negative contributor offers a specific target for faculty development. This high degree of educational interpretability increases confidence in the model's recommendations and facilitates its incorporation into real-world teaching quality improvement cycles.

Comparative Analysis with Existing Models

We compared our model against BERT-base-Chinese and optimized traditional ML baselines. BERT-base-Chinese (unfine-tuned) suffered from extreme class imbalance, achieving only 31.6% accuracy despite 81.9% weighted precision. Among traditional methods, Logistic Regression performed best (75.8% accuracy, 81.7% precision, 65.8% F1). Our model surpassed all baselines with 84.0% accuracy and 76.4% F1, while uniquely achieving 95.6% recall—critical for comprehensive screening of high-quality feedback. Compared to Logistic Regression, our model improved recall by 19.8 percentage points (95.6% vs 75.8%) with only modest precision reduction (63.6% vs 81.7%), representing a favorable trade-off for educational quality assurance where missing valuable feedback is costlier than false positives (Table 5).

Expert Consistency Assessment

Cohen's kappa coefficient was used to evaluate consistency. The kappa between Experts A and B was 0.940, indicating near-perfect agreement. Comparing Experts A and B with Expert C separately yielded kappa values of 0.965 and 0.975, respectively, both reaching near-perfect levels.

https://doi.org/10.2147/AMEP.S599543

Table 5 Comparative Analysis with Existing Models

Model	Type	Accuracy Rate	Precision	Recall	F1 Score
This Model	Custom model	0.840	0.636	0.956	0.764
BERT-base-Chinese	Deep learning	0.316	0.819 (weighted)	0.316 (weighted)	0.232 (weighted)
LR	Traditional ML	0.758	0.817	0.758	0.658
RF	Traditional ML	0.753	0.694	0.753	0.664
SVM	Traditional ML	0.753	0.692	0.753	0.655
GBM	Traditional ML	0.753	0.694	0.753	0.664

Calibration Performance

The calibration plot demonstrated good overall calibration of the LR model (Brier score: 0.0305, <0.05 indicating good reliability, Figure 8). Using quantile-based binning with equal sample sizes per bin (n≈59), the model showed acceptable reliability across most probability ranges. However, a notable deviation was observed in the high-probability region: predicted probabilities of 0.50 corresponded to observed positive fractions of 1.0, indicating mild overconfidence in this range. Conversely, at low predicted probabilities (0.02–0.12), the model exhibited slight underconfidence before converging with perfect calibration at approximately 0.15.

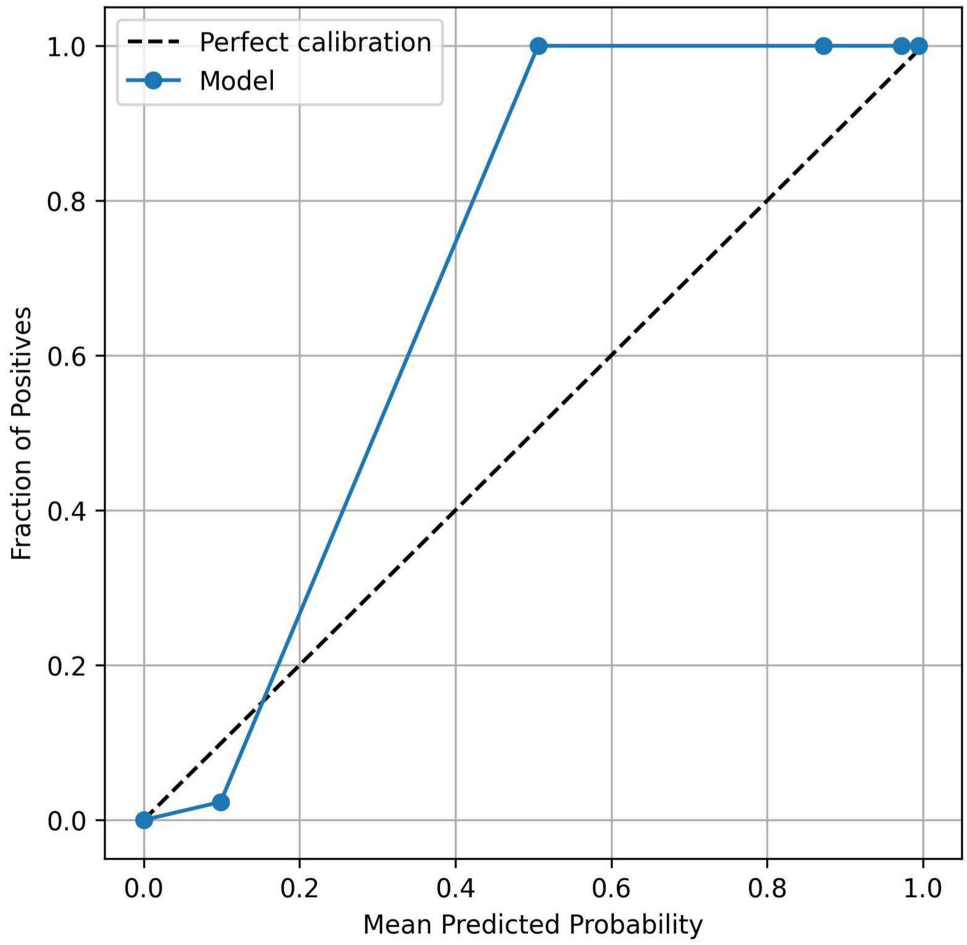

Figure 8 Calibration curve of the logistic regression model on the external test set.

https://doi.org/10.2147/AMEP.S599543

Discussion

The common challenges currently faced in evaluating medical narrative feedback in both Chinese and English mainly arise from the unique characteristics of medical texts, the complexity of language, and the subjectivity of evaluation criteria.[1] When constructing feedback evaluation models, resources and tools supporting Chinese text processing are significantly scarcer than those for English. This disadvantage is prevalent in the field of medical NLP. A recent review highlights that the application of multilingual clinical NLP is still limited by insufficient data and preprocessing challenges specific to each language. Chinese, as a non-English "high-resource language," similarly faces issues like scarce annotated corpora. This "low-resource within-domain" situation directly constrains the performance ceiling and external validity of ML models on Chinese medical education texts (such as residency training feedback).[15] This study developed and validated a machine learning-based model for automated feedback quality assessment using a specialized Chinese word segmenter and anesthesia terminology corpus. The model demonstrated strong internal performance (AUC: 0.953) and maintained acceptable discriminative ability in external validation (recall: 0.956, AUC: 0.729). The high recall supports its potential to assist teaching administrators in identifying high-quality feedback for review, though the moderate precision indicates that human oversight remains necessary. These findings suggest the feasibility of machine learning-based approaches for Chinese medical education texts, consistent with previous research.[10,16,17]

The model's combination of high recall (0.956) and moderate precision (0.636) supports its use as a preliminary screening tool. In this workflow, the model identifies potential high-quality feedback for subsequent expert review, with high recall minimizing false negatives and manual review compensating for moderate precision. This approach has the potential to streamline the evaluation process and may reduce expert workload by prioritizing feedback for review, though the actual efficiency gains require further prospective evaluation in real-world settings.

In anesthesiology residency training, high-quality feedback is recognized as important for advancing trainees' skills and may contribute to patient safety and clinical decision-making quality.[6] This model offers potential for teaching administrators to systematically monitor feedback quality and identify patterns. For instance, the model's identification of a "high proportion of vague terms" can serve as a training priority, prompting instructors to avoid generalizations and enhance feedback specificity and relevance. If integrated into teaching management systems, such tools could support data-informed faculty development and potentially facilitate quality improvement cycles, though these applications represent hypothetical benefits that require prospective implementation studies to validate.

Calibration and Clinical Utility

Despite the observed deviation in high-probability ranges, the overall Brier score of 0.0305 indicates satisfactory probability calibration for clinical application. The model's behavior in the >0.5 range creates a practical "high-confidence zone": feedback entries with predicted probabilities above 0.5 are virtually certain to be high-quality (observed positive fraction = 1.0). This supports a tiered screening strategy: Tier 1 (Threshold $\geq$ 0.24): Captures 95.6% of high-quality feedback for comprehensive review; Tier 2 (Threshold > 0.50): Identifies a subset with high precision, as evidenced by the calibration curve showing 100% actual positive rate for high-probability predictions. The excellent Brier score suggests that the model's probability outputs can be relied upon for risk stratification without additional calibration techniques in the current implementation.

Limitations

However, this study also has several limitations. First, the sample primarily originated from two hospitals. Although statistically estimated, the sample size remains relatively limited and does not cover multi-regional, multi-tiered medical institutions, which may affect the model's broad applicability. Second, manual annotation, serving as the "gold standard," retains subjective elements. Although high inter-expert agreement was achieved, the exclusion of metadata such as teaching contexts and trainee levels may overlook contextual factors influencing feedback quality. Crucially, external testing revealed the model's sensitivity to transitions in data distribution: when the proportion of high-quality data in the test set significantly exceeded that in the training set, the model tended to predict "high quality" more liberally, causing

lower precision. This suggests that site-specific calibration or threshold revisions are essential when applying the model throughout various institutions.

Suggestions for Further Research

Future research can investigate more thoroughly in the following areas: (1) Expand data sources through multi-center collaborations to collect more representative feedback data from anesthesiology residency training programs, consequently increasing the model's generalization capability; (2) Integrate deep learning models (such as BERT, ERNIE, etc.) for deep semantic mining. Combine these with the feature engineering approach from this study to build combined models that balance explainability and performance; (3) Explore model calibration and self-adjusting learning mechanisms to accommodate variations in feedback data spread throughout institutions and time periods; (4) Expand evaluation aspects beyond merely "high-quality" assessments by further categorizing feedback types (eg., skill-based, attitude-based, communication-based feedback) to provide more granular teaching insights.

Conclusion and Recommendations

This study developed and externally validated a machine learning model based on logistic regression for evaluating narrative feedback quality in anesthesiology residency training. The model demonstrated high recall (0.956) in external validation, supporting its use as a preliminary screening tool to assist teaching administrators in prioritizing feedback for manual review. With moderate discriminative ability (AUC: 0.729), the model is intended as a decision support aid rather than an independent evaluator. Site-specific calibration is recommended for cross-institutional application. This study confirms the feasibility of machine learning-based Chinese feedback quality evaluation to improve residency education management efficiency while maintaining essential human oversight.

Abbreviations

CBME, competency-based medical education; ACGME, Accreditation Council for Graduate Medical Education; EPAs, Entrustable Professional Activities; AUC-ROC, area under the receiver operating characteristic curve; ML, machine learning; NLP, natural language processing; LR, logistic regression; RF, random forests; GBM, gradient boosting machine; SMOTE, synthetic minority oversampling technique; TF-IDF, term frequency-inverse document frequency; CI, confidence interval; SHAP, Shapley additive explanations; GUI, graphical user interface; SQUIRE-EDU, Standards for Quality Improvement Reporting Excellence in Education; Mini-CEX, mini clinical evaluation exercise; DOPS, direct observation of procedural skills.

Data Sharing Statement

To ensure reproducibility and transparency, the anonymized dataset generated and analyzed during this study has been deposited in the Science Data Bank (ScienceDB) repository and is publicly available under the accession DOI: 10.57760/sciencedb.34742. The primary analysis code is publicly available on GitHub at: https://github.com/ylf3334/-The-ML-based-feedback-quality-analysis-system-for-anesthesiology-residency-training (archived version at DOI: 10.57760/sciencedb.34742). The data is released under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, and the code under an MIT License. For any questions or if you require further assistance regarding the dataset, the corresponding author is available to help.

Ethics Approval Statement

The research was carried out following the principles of the Declaration of Helsinki. The Ethics Committee approved this study, granting an exemption from review and waiving the need for informed consent (Approval No.: Ningbo University Affiliated First Hospital Ethics Review 2025 Research No. 321A-01).

Acknowledgments

We express our heartfelt gratitude to Xu Xia, Director of Teaching at the Anesthesiology Residency Training Base of Li Huili Hospital in Ningbo, Zhejiang Province, China, for her support in external testing. During the manuscript

preparation, the authors used Grammarly (v1.2.227.1811) for grammar and spelling checks only. All intellectual content remains the sole responsibility of the authors.

This study was developed and reported in accordance with the SQUIRE-EDU (Standards for Quality Improvement Reporting Excellence in Education) guidelines. A completed SQUIRE-EDU checklist is provided as Supplementary Table 1.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This research was supported by the following grants: Zhejiang Provincial Medical and Health Science and Technology Program (No. 2025KY1333); Ningbo Top Medical and Health Research Program (No. 2024010317); The First Affiliated Hospital of Ningbo University Educational Research Project (No. 2025-JXK-007).

Disclosure

No potential conflict of interest was reported by the authors. The authors alone are responsible for the content and writing of this article.

References

1. Khan SA, Taiyara J, Zary N, et al. Artificial intelligence in narrative feedback analysis for competency-based medical education: a review. *Stud Health Technol Inform*. 2025;327:1423–1427. doi:10.3233/SHTI250637
2. Mitchell JD, Jones SB. Faculty Development in Feedback Provision. *Int Anesthesiol Clin*. 2016;54(3):54–65. doi:10.1097/AIA.0000000000000109
3. Neves SE, Chen MJ, Ku CM, et al. Using machine learning to evaluate attending feedback on resident performance. *Anesth Analg*. 2021;132(2):545–555. doi:10.1213/ANE.0000000000005265
4. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Med Teach*. 2007;29(9):855–871. doi:10.1080/01421590701775453
5. Ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad Med*. 2007;82(6):542–547. doi:10.1097/ACM.0b013e31805559c7
6. Weller J, Gotian R. Evolution of the feedback conversation in anaesthesia education: a narrative review. *Br J Anaesth*. 2023;131(3):503–509. doi:10.1016/j.bja.2023.05.021
7. Choo EK, Woods R, Walker ME, et al. The quality of assessment for learning score for evaluating written feedback in anesthesiology postgraduate medical education: a generalizability and decision study. *Can Med Educ J*. 2023;14(6):78–85. doi:10.36834/cmej.75876
8. Maimone C, Dolan BM, Green MM, et al. Utilizing natural language processing of narrative feedback to develop a predictive model of pre-clerkship performance: lessons learned. *Perspect Med Educ*. 2023;12(1):141–148. doi:10.5334/pme.40
9. Solano QP, Hayward L, Chopra Z, et al. Natural language processing and assessment of resident feedback quality. *J Surg Educ*. 2021;78(6):e72–e77. doi:10.1016/j.jsurg.2021.05.012
10. Yilmaz Y, Jurado Nunez A, Ariaeinejad A, et al. Harnessing natural language processing to support decisions around workplace-based assessment: machine learning study of competency-based medical education. *JMIR Med Educ*. 2022;8(2):e30537. doi:10.2196/30537
11. Spadafore M, Yilmaz Y, Rally V, et al. Using natural language processing to evaluate the quality of supervisor narrative comments in competency-based medical education. *Acad Med*. 2024;99(5):534–540. doi:10.1097/ACM.0000000000005634
12. Sun S, Hu Q, Xu F, et al. Medical named entity recognition based on domain knowledge and position encoding. *BMC Med Inform Decis Mak*. 2025;25(1):235. doi:10.1186/s12911-025-03037-0
13. Zong H, Wu R, Cha J, et al. Advancing Chinese biomedical text mining with community challenges. *J Biomed Inform*. 2024;157:104716. doi:10.1016/j.jbi.2024.104716
14. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441. doi:10.1136/bmj.m441
15. Elvas LB, Almeida A, Ferreira JC. Natural language processing in medical text processing: a scoping literature review. *Int J Med Inform*. 2025;204:106049. doi:10.1016/j.ijmedinf.2025.106049
16. Nogueira R, Eguchi M, Kasmirski J, et al. Machine learning, deep learning, artificial intelligence and aesthetic plastic surgery: a qualitative systematic review. *Aesthetic Plast Surg*. 2025;49(1):389–399. doi:10.1007/s00266-024-04421-3
17. Verghese BG, Iyer C, Borse T, et al. Modern artificial intelligence and large language models in graduate medical education: a scoping review of attitudes, applications & practice. *BMC Med Educ*. 2025;25(1):730. doi:10.1186/s12909-025-07321-5