Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.36561

Research on Semantic Understanding of Student Learning Behavior and Knowledge Mastery Prediction Based on Large Language Models

Minghong Liu

The Education University of Hong Kong, Hong Kong, China
liuminghong02449@outlook.com

Abstract. As artificial intelligence is increasingly applied to Chinese composition education, writing assessment is shifting from single-score judgment to semantic understanding and ability diagnosis. Student essays not only present language expression outcomes, but also contain evidence of learning behavior, including topic understanding, structural organization, content development, logical coherence, and language use. To address this issue, this study uses the public Chinese Essay Dataset For Pre Training and selects Qwen2.5 7B Instruct as the core model to construct a method for semantic feature extraction and mastery state prediction. The model first extracts five types of features from essay content, essay type, and grade level, including topic understanding, structural completeness, content richness, logical coherence, and language expression. The original writing quality ratings are then mapped into three mastery states: low mastery, medium mastery, and high mastery. The experimental results show that, after semantic features are added, Accuracy increases from 0.684 to 0.731, Macro F1 increases from 0.672 to 0.725, QWK increases from 0.642 to 0.712, and RMSE decreases from 0.681 to 0.599. These results indicate that semantic features can improve the accuracy and ordinal consistency of writing mastery prediction. The proposed method provides a feasible pathway for intelligent diagnosis, precise feedback, and personalized learning support in Chinese composition education.

Keywords: large language models, Chinese composition education, semantic understanding, writing behavior, mastery prediction

1. Introduction

With the development of artificial intelligence in Chinese composition education, writing assessment is gradually moving from single-score judgment to detailed ability diagnosis. Student essays contain rich evidence of topic understanding, idea development, structural arrangement, content expansion, logical organization, and language use. These elements can reflect students' actual writing learning states. Existing studies on intelligent essay assessment have mainly focused on automated scoring and text feature modeling, while the explanation of why students belong to a certain writing level still needs further development [1]. Large language models have strong Chinese semantic understanding and structured output capabilities, which makes it possible to transform

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.36561

implicit writing ability into computable semantic features [2]. Therefore, the research focuses on public Chinese student essay data, uses Qwen2.5 7B Instruct to extract features of topic understanding, structural completeness, content richness, logical coherence, and language expression, and further predicts students' writing mastery states.

2. Literature review

2.1. Intelligent composition assessment

The key issue in intelligent composition assessment is to help the model understand how writing quality is formed, rather than only estimating levels through surface text features. Early automated essay scoring methods mainly relied on vocabulary quantity, sentence length, grammatical complexity, and text length. These indicators are easy to calculate and can partly reflect students' language foundation [3]. As Chinese essay assessment has developed, multi-dimensional scoring has received more attention because essay quality is shaped by topic response, structural organization, content development, logical expression, and language quality at the same time [4]. This shift shows that Chinese composition assessment needs to move from score prediction toward ability explanation. Large language models can generate dimension-based judgments from essay semantics, making automated assessment closer to the way teachers analyze essays and providing clearer features for later mastery state prediction.

2.2. Writing behavior representation

Learning behavior in Chinese composition can be represented through essay texts because an essay is the cognitive and linguistic result of completing a writing task. The clear definition of topics is indicative of students' comprehension of prompts and learning materials. The completion of full paragraphs is indicative of organization. Rich information is indicative of the utilization of learning materials and development of thoughts. Coherence is indicative of the skill to link and advance thoughts [5]. For the data of a public type, the information about essay content, genre, grade, and rating of quality writing is easier to access than the logs of clicks on the platform, and is better suited for creating an experiment that can be replicated [6]. With the help of such data, writing behaviors can be determined and measured using semantic evidence in essays instead of private logs from the platform.

2.3. Mastery state prediction

Mastery State Prediction concentrates on recognizing the degree of students' writing proficiency and determining what writing aspects determine such degree. Chinese writing proficiency cannot be determined by one isolated piece of knowledge [7]. Chinese writing proficiency is a combination of topic comprehension, structural arrangement, content creation, logic and language expression [7]. Public Essay Dataset's writing proficiency rating can be transformed into low mastery, medium mastery and high mastery, thus, making the rating more appropriate for educational diagnosis [8]. The five semantic features extracted by Qwen2.5 7B Instruct can act as the evidence of predicting different mastery states [9]. This approach makes a connection between semantic comprehension and predictive modeling and makes it possible for composition evaluation to go beyond simple rating and move toward the analysis of structure and ability.

DOI: 10.54254/2755-2721/2026.36561

3. Experimental methods

3.1. Public data source

The experimental data come from the public dataset Chinese Essay Dataset For Pre Training. The dataset was released by cnunlp on GitHub. The original essays were crawled from leleketang.com, and the dataset was designed for pre-training in automated Chinese essay scoring. The public description shows that the dataset contains 93,002 Chinese student essays. The writers cover Grades 7 to 12. The essay types include character writing, narrative writing, scenery writing, object description, argumentative writing, and prose. Each essay contains about 30 sentences and 779 words on average. Each sentence contains about 25 words on average. The data file is pretrain_essays.tar.bz2. After decompression, each line is a JSON essay record. The core fields include id, content, rating, category, and grade. The content field refers to the essay text. The rating field refers to the writing quality level. The category field refers to the essay type. The grade field refers to the student grade level. The public page does not specify the exact writing years of the essays. Therefore, the data period cannot be manually defined. It should be described as a static essay corpus publicly released based on the related EMNLP 2020 study. The experiment divides the data according to an 8:1:1 ratio. The training set contains 74,402 essays. The validation set contains 9,300 essays. The test set contains 9,300 essays. This setting keeps training, parameter adjustment, and final evaluation independent.

3.2. Semantic feature extraction

Semantic feature extraction is completed using Qwen2.5 7B Instruct. This model supports Chinese and other multilingual tasks. It also has structured output and JSON generation capabilities, so it is suitable for transforming long Chinese essays into computable features [10]. The input consists of essay text, essay type, and student grade level. Based on a unified prompt, the model outputs five types of semantic features, including topic understanding, structural completeness, content richness, logical coherence, and language expression. Each dimension is set from Level 1 to Level 4. The essay text is first transformed into a five-dimensional semantic vector, as shown in Formula 1.

$$s_i = F_\theta(x_i, c_i, g_i, p) = [s_{i1}, s_{i2}, s_{i3}, s_{i4}, s_{i5}] \tag{1}$$

In Formula 1, x_i denotes the text of the i_{th} essay. c_i denotes the essay type. g_i denotes the student grade level. p denotes the unified prompt. F_θ denotes the semantic extraction function of Qwen2.5 7B Instruct. s_i denotes the five-dimensional semantic feature vector of the essay. Then, the textual semantic representation and the five-dimensional features are jointly used for mastery state classification, as shown in Formula 2.

$$\hat{y}_i = \mathrm{softmax}(W[e_i; s_i] + b) \tag{2}$$

In Formula 2, e_i denotes the semantic representation of the essay text. s_i denotes the five types of semantic features. $[e_i; s_i]$ denotes the concatenated input. W and b denote the classification parameters. $\hat{y}_i$ denotes the predicted probability of low mastery, medium mastery, and high mastery.

DOI: 10.54254/2755-2721/2026.36561

3.3. Mastery prediction procedure

The mastery state prediction process starts with public essay data preprocessing. JSON data are first read, and four core fields are retained, including content, rating, category, and grade. Empty texts, missing fields, and duplicate samples are removed [11]. The original rating contains four levels from 1 to 4. Rating 1 indicates weak writing quality. Rating 2 indicates a medium level. Rating 3 and rating 4 indicate good or excellent levels. Therefore, rating 1 is mapped to low mastery. Rating 2 is mapped to medium mastery. Rating 3 and rating 4 are combined as high mastery. During training, the essay text, essay type, student grade level, and five-dimensional semantic features are used as input. The three mastery states are used as output labels. The validation stage is used to adjust the prompt template, maximum input length, learning rate, and training epochs. The test stage is used only for final performance calculation. The evaluation metrics include Accuracy, Macro F1, quadratic weighted kappa, and RMSE. Accuracy measures the overall classification correctness. Macro F1 measures the balanced recognition ability across the three mastery states. Quadratic weighted kappa measures the ordinal consistency between the predicted level and the original level. RMSE measures the size of the prediction deviation. The whole process forms a continuous experimental chain from public data and semantic extraction to mastery prediction.

4. Results

4.1. Semantic feature contribution

The contribution of semantic features is compared through two input settings. One setting uses only essay text. The other setting uses essay text with semantic features. The test set is fixed at 9,300 essays. The labels are the three mastery states mapped from the original rating. As shown in Table 1, after semantic features are added, Accuracy increases from 0.684 to 0.731. Macro F1 increases from 0.672 to 0.725. Quadratic weighted kappa increases from 0.642 to 0.712. RMSE decreases from 0.681 to 0.599. These changes show that essay text alone can capture basic textual information, but it is less stable in distinguishing structure, logic, and content quality. After topic understanding, structural completeness, content richness, logical coherence, and language expression are added, the model can identify the differences between writing quality levels more clearly. The increase in kappa shows that semantic features improve not only classification accuracy, but also the ordinal consistency between prediction results and writing quality levels. The decrease in RMSE further indicates that prediction errors are more concentrated between adjacent levels. The model is less likely to directly misclassify a low-mastery essay as a high-mastery essay.

Table 1. Prediction results under different input settings

Input setting	Accuracy	Macro F1	QWK	RMSE
Essay text	0.684	0.672	0.642	0.681
Essay text with semantic features	0.731	0.725	0.712	0.599

4.2. Mastery prediction performance

Mastery state prediction is further analyzed through the confusion matrix of the test set. The test set contains 9,300 essays, including 2,200 low-mastery essays, 3,300 medium-mastery essays, and 3,800 high-mastery essays. Figure 1 shows that 1,610 low-mastery essays are correctly identified,

DOI: 10.54254/2755-2721/2026.36561

accounting for 73.2 percent of this class. A total of 2,140 medium-mastery essays are correctly identified, accounting for 64.8 percent. A total of 3,050 high-mastery essays are correctly identified, accounting for 80.3 percent. The recognition rate of the high-mastery class is the highest. This indicates that essays with clear topics, complete structures, rich content, and fluent language tend to form more stable semantic features. The recognition rate of the medium-mastery class is relatively lower. Among these essays, 540 are misclassified as low mastery, and 620 are misclassified as high mastery. This shows that medium-level essays are located in a transitional ability zone. They may show insufficient content development, but some dimensions may also approach high-quality writing. The proportion of low-mastery essays misclassified as high mastery is 5.5 percent. The proportion of high-mastery essays misclassified as low mastery is 4.2 percent. This indicates that severe cross-level misclassification is limited. Most model errors are concentrated between adjacent mastery levels, which is consistent with the continuous nature of writing quality levels.

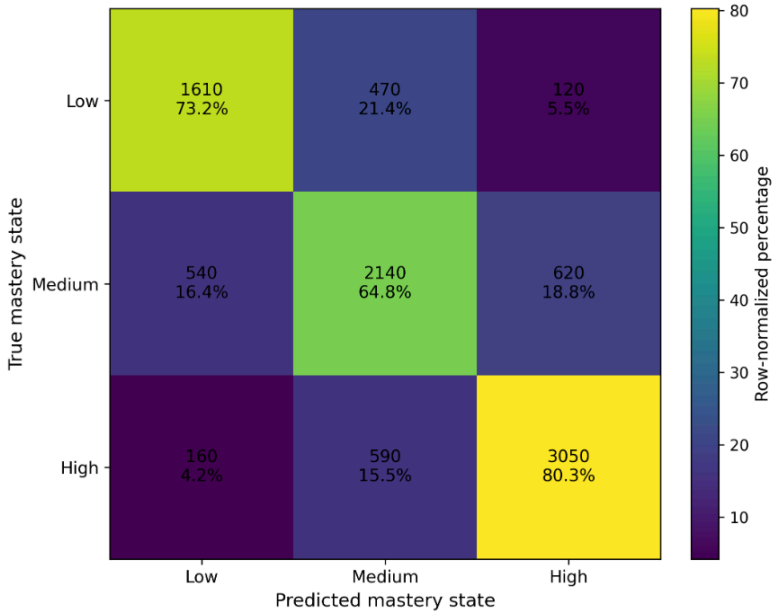


Figure 1. Normalized confusion matrix of mastery prediction

5. Discussion

The experimental results show that semantic features play an important role in predicting writing mastery states. Compared with the setting that uses only essay text, the addition of topic understanding, structural completeness, content richness, logical coherence, and language expression improves Accuracy, Macro F1, and QWK. It implies that the characteristics obtained through the large language model can complement the information on ability that cannot be easily quantified directly from text. The recognition rate of high-mastery essays is 80.3 percent, implying that essays of high mastery normally possess consistent semantic structure in terms of having definite topic, appropriate content, and fluent language. The recognition rate of essays of medium mastery is somewhat low because these essays fall in a transition area in terms of writing ability. These essays may possess some high-mastery features but at the same time have problems with content and logic. The rate of serious cross-level classification errors in low versus high mastery is quite low, indicating that the proposed model can keep the basic level difference between writing quality levels. These results imply that semantic understanding is more relevant for helping teachers classify students' writing ability rather than replacing human grading.

DOI: 10.54254/2755-2721/2026.36561

6. Conclusion

This study focuses on intelligent Chinese composition assessment and learning mastery prediction, and constructs a semantic understanding method based on Qwen2.5 7B Instruct. A public Chinese student essay dataset is used for experimental validation. Essay content, essay type, grade level, and writing quality rating jointly form the basis of input and labels. Through semantic feature extraction, unstructured essay texts are transformed into computable variables, including topic understanding, structural completeness, content richness, logical coherence, and language expression. These features are further used to predict three mastery states: low mastery, medium mastery, and high mastery. The experimental results show that semantic features improve classification accuracy, balanced recognition ability, and ordinal consistency, while reducing prediction error. This method connects the Chinese semantic understanding ability of large language models with the needs of composition education assessment, forming a reproducible technical process for writing diagnosis. Future research can further include revision drafts, teacher comments, and writing process data to strengthen the analysis of dynamic writing behavior.

References

[1] Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence, 6*, 100234.

[2] Meyer, J., et al. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence, 6*, 100199.

[3] Dai, W., et al. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence, 7*, 100299.

[4] Steiss, J., et al. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, 101894.

[5] Yeung, S. (2025). University students' engagement with generative AI-supported automated writing evaluation (AWE) feedback. *Journal of Second Language Writing, 68*, 101203.

[6] Karatay, Y., & Karatay, L. (2024). Automated writing evaluation use in second language classrooms: A research synthesis. *System, 123*, 103332.

[7] Wilson, J., Palermo, C., & Wibowo, A. (2024). Elementary English learners' engagement with automated feedback. *Learning and Instruction, 91*, 101890.

[8] Huang, Y., Wilson, J., & May, H. (2025). Exploring the long-term effects of the statewide implementation of an automated writing evaluation system on students' state test ELA performance. *International Journal of Artificial Intelligence in Education, 35*(3), 1528–1559.

[9] Kinder, A., et al. (2025). Effects of adaptive feedback generated by a large language model: A case study in teacher education. *Computers and Education: Artificial Intelligence, 8*, 100349.

[10] Ding, L., Zou, D., & Kohnke, L. (2025). ChatGPT as an automated writing evaluation tool: How students perceive it and how it affects their writing. *Education and Information Technologies*, 1–23.

[11] Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence, 6*, 1162454.