ELion: An Intelligent Chinese Composition Tutoring System Based on Large Language Models

Chanjin Zheng  
*East China Normal University*

Shaoyang Guo  
*East China Normal University*

Wei Xia  
*East China Normal University*

Shaoguang Mao  
*Microsoft Research Asia*

Follow this and additional works at: https://www.ce-jeme.org/journal

**Recommended Citation**

Zheng, Chanjin; Guo, Shaoyang; Xia, Wei; and Mao, Shaoguang (2023) "ELion: An Intelligent Chinese Composition Tutoring System Based on Large Language Models," *Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语期刊*: Vol. 4: Iss. 3, Article 3.  
DOI: https://doi.org/10.59863/MPJO6480  
Available at: https://www.ce-jeme.org/journal/vol4/iss3/3
ELion: An Intelligent Chinese Composition Tutoring System Based on Large Language Models

Chanjin Zheng a b, Shaoyang Guo a, Wei Xia b and Shaoguang Mao c

a Shanghai Institute of the Artificial Intelligence for Education, East China Normal University
b Department of Educational Psychology, the Faculty of Education, East China Normal University
c Microsoft Research Asia

Abstract
For a long time, Chinese language teachers in primary and secondary schools have been confronting challenges of heavy workload, low efficiency, and difficulty in improving the quality of composition evaluations. This article introduces “ELion”, an intelligent Chinese composition tutoring system based on large language models. The system utilizes deep linguistic features to evaluate the quality of compositions and provide interpretable feedback. By discussing the overall design, evaluation framework structure, and scoring algorithm principles of ELion, this paper addresses the theoretical, technical, and engineering issues of intelligent evaluation of Chinese compositions in the educational context. Small-scale experiments conducted in schools demonstrate that ELion performs well in language error detection, rhetorical techniques, and the expression of actions and emotions. It can basically meet the needs of Chinese language teaching in primary and secondary schools. In the future, ELion will further develop algorithms for “instruction-learning-evaluation” alignment assessment, and personalized precise feedback generation, based on the GPT model. This will improve the evaluation effectiveness in topic analysis, text structure, and genuine emotional expression. Additionally, systematic field experiments for the system will be conducted to explore the application of artificial intelligence in education.

Keywords
ELion Intelligent Chinese Composition Tutoring System; Large Language Models; Intelligent Evaluation; Chinese Composition Assessment

1 Introduction

Language is the most essential tool for communication in human society. The cultivation and application of language proficiency accompany individuals throughout their development stages. Among various language skills, writing is not only a central aspect of language learning but also one of the most advanced and most complex language abilities, involving multiple elements such as memorization, understanding, applying, and expressing ideas through written words. Developing individual writing skills has always been a challenge and pain point in language learning: the difficulty lies in the unavoidable fact that improving writing skills requires targeted feedback and multiple rounds of revisions, while the pain lies in the high costs of time and effort required for teachers to grade and for students to revise compositions, resulting in limited improvements in students’ writing abilities in educational practice. Therefore, reducing the cost of composition assessment and revision in the process of cultivating writing skills has become a central concern in academia.

As shown in Figure 1, the ELion Intelligent Chinese Composition Tutoring System (https://elion-student.aiecnu.cn; experience code: dc67fc90) is a fourth-generation automated Chinese composition scoring system. This system is jointly developed by the Shanghai Institute of Artificial Intelligence for Education at East China Normal University and Microsoft Research Asia. It addresses the theoretical, technical, and engineering challenges of intelligent evaluation of Chinese
compositions in an educational context using large language models (LLMs). This paper first reviews the development process of automated essay scoring systems from the perspective of natural language processing techniques. Then, it provides a comprehensive introduction to the overall design, evaluation framework, scoring algorithms, as well as the system’s implementation and feedback in schools. Finally, considering the breakthroughs in generative artificial intelligence technology, the paper discusses several future directions for intelligent evaluation of Chinese compositions.

Figure 1: Interface of ELion Intelligent Composition Tutoring System

2 Overview of the Development of Automated Essay Scoring Systems from the Perspective of Natural Language Processing (Xia et al., 2023)

Writing practice plays a crucial role in the process of language learning. However, due to the large number of students in China, teachers are facing an increasingly heavy burden when it comes to grading student essays. Therefore, Automated Essay Scoring (AES) systems have gained widespread attention both domestically and internationally. AES technology utilizes computer algorithms to replace human scoring, aiming to achieve automatic evaluation of essays. With the assurance of scientifically sound scoring algorithms, AES significantly improves grading efficiency, reduces evaluation errors caused by subjective fluctuations among graders, and also lowers the cost of manual grading. Furthermore, the appropriate use of AES systems can help students better understand key points in writing and enhance their writing abilities.

The underlying technology applied to AES systems, Natural Language Processing (NLP), encompasses a wide range of fields, including disciplines such as linguistics and computer science. From the perspective of linguistic units, NLP includes different granularities such as characters, words, sentences, paragraphs, and discourse. From the perspective of task types, it covers various categories including processing, understanding, and generation. The following outlines the preliminary categorization of NLP tasks, which, in turns, underlies the development stages of automated essay scoring systems.

2.1 1950 to 1990: Rule-based Research Stage

During the rule-based research stage, linguistic knowledge was incorporated into automated essay scoring systems by formulating distinct rules associated with vocabulary and formal grammar. As shown in Figure 2, this stage was comprised of four main components: data construction, rule formulation, rule application, and effectiveness evaluation. The key of this stage was to define the form of rules, aiming to enable linguists to easily convert their knowledge into rules, even if they might lack knowledge of computer programming.

During this stage, the first automated essay scoring system called Project Essay Grade (Page, 1966), was developed.
This system primarily utilized feature rules to represent the fluency of essays and achieve automated scoring. By extracting relevant rules from a set of training essays, the rules were transformed and used to establish a regression model for predicting essay grades. Ultimately, the corresponding essay scores were generated through the regression equation. However, due to the limitations of computing technology at that time and the lack of widespread interest in automated essay scoring, there was no significant progress in AES for over two decades following the development of Project Essay Grade.

![Figure 2: Basic Process of Rule-based Natural Language Processing](image)

2.2 1990 to 2010: Feature-based Statistical Machine Learning Research Stage

Since 1990, AES has attracted growing interest and attention with the increasing popularity of computer technology. In the late 1990s, a series of AES systems such as the Intelligent Essay Assessor (Landauer et al., 1998) and E-raters (Attali & Burstein, 2006) were released one after another. These systems marked the advent of the feature-based statistical machine learning research stage.

During this stage, most natural language processing algorithms implemented supervised classification to convert natural language processing tasks into classification tasks. These tasks involved constructing feature representations based on the task’s characteristics and training models using large-scale labeled corpora. As depicted in Figure 3, this process primarily consisted of four steps: data construction, data preprocessing, feature construction, and model learning. The key aspect of this approach was artificial feature construction, utilizing fundamental natural language processing algorithms to preprocess the raw data, selecting suitable machine learning models, determining learning criteria, and applying corresponding optimization algorithms.

In this stage, AES systems predominantly utilized corpus-based modeling approaches, employing real essay data for analysis. Statistical machine learning techniques were applied to preprocess the essay data, which included analyzing vocabulary usage, discourse structure, and semantics. Based on the analysis results, features were extracted to construct predictive factors for selecting and weighting essay scores. Ultimately, the final scores were computed using machine learning models and evaluated against benchmark sets of human-scored essays.

2.3 2010 to 2018: Deep Neural Network Research Stage

Due to the limitations of feature-based statistical machine learning, which only extracts predetermined features and has limited evaluation criteria, a significant gap existed between the extracted features and the dimensions and depth that teachers focus on when evaluating essays. Therefore, as shown in Figure 4, the research stage based on deep neural
networks aimed to address these limitations by constructing models with certain “depth” that integrate feature learning and prediction models. The goal was to optimize the algorithms so that the models could automatically learn superior feature representations and make predictions based on them. This stage effectively resolved the drawbacks of the previous stage in terms of feature extraction for evaluation metrics and alignment with real-world scenarios. Typically, this stage includes three parts: data construction, data preprocessing, and model learning. The core of this approach is to transform the raw data into more abstract representations through multi-layer feature transformations, enabling the learned representations to partially replace manually designed features. The key aspect is representation learning, which involves constructing multi-layer feature representations with certain depth.

Taghipour and Ng (2016) were the first to use Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to construct AES systems. They utilized these two neural networks to learn vector representations of essays and employed fully connected layers to score the essays. Following this, Dong and Zhang (2016) further explored the use of word-level CNN+Pooling and sentence-level CNN+Pooling to encode sentences and the entire essay, respectively. Their goal was to model the hierarchical structure of an essay. In further research, Dong and Zhang (2016)
introduced attention mechanisms to differentiate the importance of words and sentences, demonstrating that CNN is more suitable for modeling sentences, while RNN is more appropriate for modeling the entire essay. However, when using RNN to model the entire essay, information transfer issues may arise when there are a large number of sentences, as the system’s state could be influenced by preceding sentences. To address this problem, Sak et al. (2014) proposed using essay features for score prediction, while Alikaniotis et al. (2016) introduced Long Short-Term Memory (LSTM) networks for feature extraction, effectively resolving the issue mentioned above.

2.4 2018 to 2022: Pretraining + Fine-tuning Research Stage

Since the introduction of the ELMo model (Peters et al., 2018) in 2018, deep learning-based natural language processing has evolved into a “pretraining + fine-tuning” research approach. As depicted in Figure 5, this approach involves pretraining the model using self-supervised tasks to learn more general language representations from large-scale corpora. Subsequently, the pretrained network is fine-tuned to address previous issues such as slow training speed, low efficiency, and overfitting in deep neural networks. This method has shown excellent performance in various natural language processing tasks. Pretraining tasks can include self-supervised or supervised approaches, such as language models, masked language models, denoising autoencoders, and contrastive learning. Additionally, extension tasks such as knowledge graphs, multilingual, and multimodal learning can also be incorporated.

![Figure 5: Basic Process of Pretraining + Fine-tuning for Natural Language Processing](image)

For Example, BERT (Bidirectional Encoder Representations from Transformers) aims to pretrain deep bidirectional text representations by jointly conditioning the context of all network layers. By adding an additional output layer, the pretrained BERT model can be fine-tuned to create optimal models applicable to a wide range of tasks without substantially modifying the architecture for specific tasks (Devlin et al., 2019). Based on this idea, Yang et al. (2020) proposed the R2BERT model (BERT model with regression and ranking), which fine-tunes the pretrained BERT to learn text representations that capture deep semantic information. A fully connected neural network is used to map these representations to scores. Finally, the scoring is jointly constrained by combining regression loss and batch-wise ranking loss.

3 Development and Testing of ELion Chinese Essay Intelligent Tutoring System

3.1 System Development Approach

Design-Based Research (DBR) is a research method employed by educational researchers to systematically design and implement a learning environment in authentic educational contexts, examining its effectiveness (Barab, 2014). In
this study, more than 20 primary and secondary schools in Shanghai, Jiangsu, and Sichuan participated in the design and implementation of writing learning and assessment activities within a computer-supported writing learning environment. Through an iterative process, the design and development of the intelligent essay evaluation system were carried out, the effect is tested through a series of educational experiments such as teaching practice. The research data and results were used to further improve the system, thus realizing the development and design of the automated composition assessment system.

To ensure that the system effectively improves the efficiency and quality of Chinese essay teaching, ELion established a team consisting of Chinese language education researchers, educational measurement experts, algorithm developers, software engineers, and implementation teams. Efficient collaboration among these teams was ensured through online real-time communication and weekly work meetings. Specifically, as shown in Figure 6, the system development of ELion followed an iterative upgrading process in five stages: user engagement (requirements elicitation), problem abstraction, theoretical summarization, technological support, and engineering research and development. In the user engagement stage, ELion formed a Chinese language education research team composed of subject experts, education researchers, and Chinese language teachers. They identified the pain points and difficulties in writing instruction through brainstorming sessions, group interviews, and needs analysis. Based on this, the educational measurement team abstracted the problems and further sought relevant theories and methods to address them, thereby determining the main functions and directions for system development. Subsequently, the algorithm development team tackled technical challenges, and the software engineering team conducted system development, resulting in the initial trial version of ELion, which was then tested in some schools by the implementation team. After a period of trial, feedback from teachers and students was collected through interviews or questionnaires. The collected feedback was then discussed by the Chinese language education research team to drive iterative upgrades of the system’s functionalities.

![Figure 6: ELion System Development Process](image)

### 3.2 System Function Design

ELion aims to provide teachers and students with a convenient writing teaching platform, advanced writing learning tools, and a comprehensive literacy assessment system. Firstly, teachers can utilize ELion for classroom instruction and post-class essay grading. Secondly, students can use ELion to identify weaknesses in their essays and engage in independent learning when teachers and parents are unavailable. Lastly, ELion can assess students’ Chinese language literacy from a writing perspective, providing a reliable tool for teaching and evaluation of Chinese language literacy.

ELion operates on a class basis, binding the teacher-student relationship, and facilitates daily teaching activities. It consists of two information-sharing subsystems: the teacher-end and the student-end, as shown in Figure 7. These subsystems enable functions such as essay assignment release, photo upload, intelligent evaluation, and multiple revisions. The teacher-side includes three main modules: class management, essay management, and essay grading. Among them,
essay grading is the core module of ELion. The teacher side provides automatic essay evaluation, modification of evaluation results, voice comments, peer evaluation, recommended exemplary essay, and request for revisions as detailed below:

**Automatic essay evaluation:** Automatically provides word, phrase, sentence, and paragraph-level comments, as well as a teacher’s overall assessment, to alleviate the grading burden for teachers.

**Modification of evaluation results:** Allow teachers to freely modify, add, or delete the results of automatic essay evaluation, ensuring a more accurate evaluation.

**Voice comments:** Allow for the recording of audio comments for an essay, enriching the content of essay evaluation and making the feedback more flexible and vivid.

**Peer evaluation:** Enable students to review and provide feedback on each other’s essays, either freely or randomly assigned, fostering students’ core competencies.

**Recommended exemplary essays:** Allow outstanding essays to be designated as exemplary and recommended for the entire class to reference and learn from.

**Request for revisions:** Enable teachers to return an essay to a student for revision based on evaluation feedback, requiring students to revise according to the feedback and resubmit the revised essay.

The student-end includes three main modules: class assignments, personal writing space, and exemplary essay collection. Class assignments correspond to teaching assignments given by teachers, personal writing space corresponds to students’ voluntary writing practice, and exemplary essay collection corresponds to reference essays recommended by teachers. The student-side focuses on the core function of intelligent essay evaluation, providing the following functions:

**Photo recognition:** Allow students to upload handwritten essays by taking photos, and the system automatically performs text recognition, reducing the difficulty of using the system.

**Text correction:** Enables students to manually correct wrongly recognized text, improving the accuracy of grading.

**Automatic evaluation reference:** Automatically provides preliminary essay modification suggestions based on eLion algorithm for uploaded essays.

**Teacher’s evaluation feedback:** Provides students with feedback on their essays from teachers, offering formal essay modification suggestions.

**Multiple revision and upload:** Supports multiple online revisions of essays, reducing the time burden for students to repeatedly rewrite their essays.

![Figure 7: Core Functions of the Teacher-side and Student-side](image)

### 3.3 Assessment Framework for Chinese Compositions

To achieve deep language feature analysis, eLion assembled a 23-member teaching and research team composed of Chinese teachers, educational researchers, evaluation experts, and AI specialists, tasked specifically with developing a composition assessment framework appropriate for the Chinese context at the primary and middle school stages. After
reviewing existing Chinese composition assessment frameworks (Song et al., 2020; Wei et al., 2022), as shown in Figure 8, the team established a four-level assessment framework that ranges from shallow to deep.

**Language application layer:** This layer primarily targets basic writing abilities, evaluating the extent to which students have mastered words, punctuation, idioms, ancient poems, and grammatical structures. It is especially suitable for lower primary school students.

**Language expression layer:** This layer mainly evaluates students’ language expression abilities from four aspects: rhetoric recognition, descriptive techniques, proficient use of words and sentences, and paragraph development, especially suitable for lower and middle primary school students.

**Discourse anomaly detection:** This layer is mainly used to judge whether students’ work contains plagiarism or instances of disorganized writing, providing validity support for the evaluation results.

**Discourse quality assessment:** This layer mainly assesses the quality of compositions from four dimensions: understanding the question, content, expression, and writing, thereby giving targeted review opinions, especially suitable for upper primary and middle school students.

### Figure 8: ELion’s Chinese Composition Assessment Framework

<table>
<thead>
<tr>
<th>Evaluation of the Quality of the Text Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>topic analysis:</strong> relevant to the topic, sound thinking, clear central idea, meets word count requirements</td>
</tr>
<tr>
<td><strong>content:</strong> standardized writing, correct punctuation, fluent language, genuine emotions</td>
</tr>
<tr>
<td><strong>expression:</strong> rigorous structure, elegant rhetoric, vivid description, smooth development</td>
</tr>
<tr>
<td><strong>writing:</strong> standardized handwriting, clear font, aesthetically pleasing characters</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Abnormality Detection in the Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>plagiarism detection</strong></td>
</tr>
<tr>
<td><strong>detection of disordered writing</strong></td>
</tr>
<tr>
<td><strong>non-healthy recognition</strong></td>
</tr>
<tr>
<td><strong>identification of a haphazard account</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Linguistic Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>rhetorical identification</strong></td>
</tr>
<tr>
<td>metaphor, parallelism, personification, quotation</td>
</tr>
<tr>
<td><strong>descriptive techniques</strong></td>
</tr>
<tr>
<td>language demeanor, appearance, scenery, psychology, actions</td>
</tr>
<tr>
<td><strong>excellent phrases and sentences</strong></td>
</tr>
<tr>
<td>beautiful sentences, advanced vocabulary</td>
</tr>
<tr>
<td><strong>paragraph expansion</strong></td>
</tr>
<tr>
<td>introduction, development, transition, conclusion, logical coherence in writing</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>correction of erroneous characters</strong></td>
</tr>
<tr>
<td><strong>grammar error detection</strong></td>
</tr>
<tr>
<td><strong>punctuation error detection</strong></td>
</tr>
<tr>
<td><strong>identification of errors in idioms/classical literature</strong></td>
</tr>
</tbody>
</table>

3.4 Large-scale Model-based Automated Essay Evaluation Algorithm

The ELion’s automated essay evaluation algorithm mainly incorporates Optical Character Recognition (OCR) technology for Chinese children’s handwriting, the intelligent evaluation technology based on the BERT model, and the intelligent comment generation technology based on predefined templates. The OCR technology for handwriting recognition is primarily used to assess the neatness of writing and the aesthetic appeal of fonts, providing a digital text foundation for subsequent analysis. The intelligent evaluation technology focuses on detecting spelling errors, inappropriate word usage, language fluency, coherence, logical reasoning, rhetorical techniques, and descriptive analysis, etc. The intelligent feedback generation technology generates personalized essay revision suggestions and comments based on the evaluation results from the previous steps, making the evaluation results more comprehensive and reliable.

Figure 9 demonstrates the operating principle of ELion’s automated essay evaluation algorithm based on large-scale
language models. After students upload photos of their essays, the algorithm first performs text recognition to convert the images into text encoding while generating basic descriptive statistics such as word count and paragraph count. Then, using the BERT model, the algorithm represents sentences and constructs a joint loss model to analyze words, phrases, sentences, paragraphs, and the overall essay. The algorithm provides grade evaluations and correction information for the essays. Finally, based on the analysis results and predefined comment templates, personalized report of evaluation results and essay comments are generated.

![Figure 9: ELion’s Intelligent Essay Scoring Algorithm](image)

It is worth mentioning that though OCR technology is relatively mature and not directly related to automated essay evaluation, in the ELion project, OCR recognition of children’s handwritten Chinese characters became the first technical and engineering problem that needed to be overcome. The reasons for this are as follows: Firstly, taking photos with smartphones is the most suitable method for inputting essays in China. In large-scale exams in our country, especially for entrance exams, traditional pen-and-paper exams still require students to write essays, and handwriting is an important criterion for essay scoring. Therefore, the system cannot require students to type on computers or use voice input, as these methods have faced strong opposition from teachers. Moreover, schools require the system to evaluate handwriting. Secondly, several specific situations in essay handwriting recognition pose challenges to OCR recognition in this project. Essay input involves a large amount of text to be recognized at once, and high accuracy in recognition is required. Spending too much time correcting OCR recognition errors would lead to a poor user experience, prolonged screen time for students, and widespread reluctance to use the system. Finally, we found that children’s handwriting, especially
for students in the third and fourth grades of primary school, has unique characteristics, and general OCR recognition technology performs poorly, with a recognition rate of less than 70%. Taking all these factors into consideration, when using our first version of the system to input an essay, students needed approximately ten minutes to correct scanning and recognition errors, which led to some complaints.

To address this issue, we collaborated with the OCR Cognitive Services team at Microsoft Research Asia and customized the Chinese children’s handwritten OCR recognition technology. We specifically collected handwriting samples from young children and made targeted modifications and training to the OCR algorithm. Additionally, we optimized the interface for input text revision during the recognition phase. Through these measures, we significantly improved the accuracy of children’s handwriting recognition and the user experience in correcting input text errors within the system.

### 3.5 Field Implementation and Feedback in Schools

As of June 2023, approximately 65 primary and secondary schools have adopted the ELion Chinese Essay Intelligent Tutoring System. ELion has accumulated a user base of 5,549 students from grades 3 to 9, along with 245 teacher users. Over 20,000 essays have been uploaded to the system. To conduct an initial assessment of ELion’s effectiveness in accurate scoring and reducing teachers’ workload, the research team conducted in-depth interviews and surveys with 42 teachers (21 from primary schools and 21 from junior high schools). The survey required the interviewed teachers to evaluate ELion’s accuracy and workload reduction based on six aspects: language error detection (such as identifying spelling mistakes), rhetorical devices (such as metaphor and personification), descriptive techniques (such as depicting actions and expressions), theme analysis (such as addressing the topic and intent), text structure (such as organizing the composition), and conveying genuine emotions (such as expressing true feelings). The evaluation was performed using a Likert scale with five levels (higher scores indicating better alignment with the criteria) to assess ELion’s accuracy and workload reduction effects.

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Very Inaccurate (%)</th>
<th>Somewhat Inaccurate (%)</th>
<th>Moderate (%)</th>
<th>Somewhat Accurate (%)</th>
<th>Very Accurate (%)</th>
<th>Accurate and Very Accurate Proportion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language Error Detection</td>
<td>7.3</td>
<td>7.3</td>
<td>14.6</td>
<td>29.3</td>
<td><strong>41.5</strong></td>
<td>70.8</td>
</tr>
<tr>
<td>Rhetorical Methods</td>
<td>9.8</td>
<td>4.9</td>
<td>14.6</td>
<td><strong>39.0</strong></td>
<td>31.7</td>
<td>70.7</td>
</tr>
<tr>
<td>Action Expression</td>
<td>7.3</td>
<td>7.3</td>
<td>17.1</td>
<td><strong>41.5</strong></td>
<td>26.8</td>
<td>68.3</td>
</tr>
<tr>
<td>Theme Analysis</td>
<td>7.3</td>
<td>14.6</td>
<td>24.4</td>
<td>24.4</td>
<td><strong>29.3</strong></td>
<td>53.7</td>
</tr>
<tr>
<td>Article Structure</td>
<td>7.3</td>
<td>12.2</td>
<td><strong>26.8</strong></td>
<td><strong>26.8</strong></td>
<td>26.8</td>
<td>53.6</td>
</tr>
<tr>
<td>Genuine Emotions</td>
<td>7.3</td>
<td>12.2</td>
<td>22.0</td>
<td><strong>36.6</strong></td>
<td>22.0</td>
<td>58.6</td>
</tr>
</tbody>
</table>

**Note:** Bold indicates the maximum value in the same dimension.
Table 2: Assessment of ELion Algorithm’s Effectiveness in Reducing Workload (N=42)

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Very Inaccurate (%)</th>
<th>Somewhat Inaccurate (%)</th>
<th>Moderate (%)</th>
<th>Somewhat Accurate (%)</th>
<th>Very Accurate (%)</th>
<th>Accurate and Very Accurate Proportion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language Error Detection</td>
<td>4.9</td>
<td>9.8</td>
<td>14.6</td>
<td>24.4</td>
<td><strong>46.3</strong></td>
<td>70.7</td>
</tr>
<tr>
<td>Rhetorical Methods</td>
<td>7.3</td>
<td>9.8</td>
<td>19.5</td>
<td>29.3</td>
<td><strong>34.1</strong></td>
<td>63.4</td>
</tr>
<tr>
<td>Action Expression</td>
<td>9.8</td>
<td>4.9</td>
<td>24.4</td>
<td><strong>31.7</strong></td>
<td>29.3</td>
<td>61</td>
</tr>
<tr>
<td>Theme Analysis</td>
<td>9.8</td>
<td>7.3</td>
<td><strong>29.3</strong></td>
<td>29.3</td>
<td>24.4</td>
<td>53.7</td>
</tr>
<tr>
<td>Article Structure</td>
<td>9.8</td>
<td>4.9</td>
<td><strong>29.3</strong></td>
<td>26.8</td>
<td><strong>29.3</strong></td>
<td>56.1</td>
</tr>
<tr>
<td>Genuine Emotions</td>
<td>7.3</td>
<td>9.8</td>
<td>26.8</td>
<td><strong>29.3</strong></td>
<td>26.8</td>
<td>56.1</td>
</tr>
</tbody>
</table>

*Note: Bold indicates the maximum value in the same dimension.*

As shown in Table 1 and Table 2, ELion excels in language error detection with high accuracy and workload reduction (over 70% of scores above level 4). It also performs well in rhetorical devices depicting actions and expressions (over 60% of scores above level 4). However, its performance in theme analysis, text structure, and conveying genuine emotions is relatively average (scores above level 4 ranging from 53% to 58%). This indicates that ELion has achieved maturity in relatively basic language usage and expression, but there is still room for improvement in deep language analysis.

4 Summary, Discussion, and Future Outlook

Since the launch of ELion version 1.0 in July 2021, the system has experienced two version iterations and has been applied for two software copyrights and patent rights, effectively addressing the system’s functional design, evaluation framework for Chinese essays, and the development of AI algorithms based on large models. Preliminary field experiments have shown that ELion has achieved good accuracy and workload reduction in language correction and rhetorical devices, but there is room for improvement in topic analysis, text structure, and emotional expression. To address these issues, ELion plans to conduct the following research: Firstly, with the explosive breakthrough in generative AI technology, the existing BERT model used by ELion is no longer the only option for large language models. In the future, research will be conducted on the “Teaching Consistency Evaluation” algorithm based on the GPT model to improve the evaluation of theme analysis, text structure, and emotional expression. Secondly, the current feedback generation technology of ELion relies mainly on preset templates, lacking personalization. Future research will focus on precise automatic comment generation and other techniques. Lastly, the field experiments of ELion have limitations in terms of sample size and research conclusions. Future efforts will involve expanding the sample size, conducting AI social experiments, and promoting educational equity in Chinese language learning abroad.

Currently, there are three aspects to consider regarding the ELion Chinese essay intelligent tutoring system: Firstly, the recent breakthroughs in generative AI have had a significant impact on the automated evaluation of Chinese writing.

Although the “pre-training + fine-tuning” research phase has achieved remarkable results in the field of natural language processing, there are still many challenges to address, such as improving the model’s robustness and interpretability. In November 2022, the release of ChatGPT sparked widespread discussion and had a significant impact on natural language processing tasks. ChatGPT is capable of completing various types of natural language processing tasks through conversation with users based on task instructions. The main process includes the construction of large-scale language models, the injection of general task capabilities, and the utilization of the model, as shown in Figure 10.
The technologies mentioned above represent a disruptive development in natural language processing. The technology of “content generation”, which was previously considered as the most challenging aspect of natural language processing, is still in its early stages of development. However, the breakthrough progress of ChatGPT has largely addressed this challenging problem. The performance of these technologies in more difficult aspects, such as “topic development and expansion” and “coherence and logic” requires further exploration. However, it is foreseeable that these technologies hold the potential to reshape the landscape of automated essay evaluation.

Secondly, the intelligent evaluation of advanced Chinese writing abilities requires further exploration.

The “overall plan for deepening education evaluation reform in the new era,” issued by the Central Committee of the Communist Party of China and the State Council in October 2020, serves as the guiding document for educational evaluation reform in China’s new era. Regarding the evaluation of primary and secondary education, the document emphasizes the key task of “adhering to scientific and effective approaches, improving outcome evaluation, strengthening process evaluation, exploring value-added evaluation, improving comprehensive evaluation, fully utilizing information technology, and enhancing the scientific, professional, and objectivity nature of educational evaluation.” Among the four forms of evaluation, process evaluation occupies a central position (Xin Tao & Jia Yu, 2019).

Furthermore, the latest versions of the “General High School Curriculum Plan (2017 Edition, Revised in 2020)” and the “Compulsory Education Curriculum Plan (2022 Edition)” (referred to as the “Curriculum Plan” below) clearly state that the basis for evaluating subject core literacy is the “Academic Quality Standards.” “The standards that subject core literacy should achieve and the key performances at each level constitute the criteria for evaluating learning quality,” which is used to directly guide educational evaluation. Academic quality standards have a phased development characteristic. For example, in the Chinese language subject, the “Compulsory Education Chinese Language Curriculum Standards (2022 Edition)” clearly states, “Academic quality and stage requirements are integrated, providing a clear depiction of an advanced level of Chinese language academic development.” Therefore, “learning progression” is the most important form of representing academic quality standards.

Learning progression provides a detailed description of the development stages of a student’s advanced and complex skills, from low to high and from single to compound. Researchers in the field of core literacy emphasize the importance of learning progression in core literacy development and evaluation. For example, learning progression is an important tool for curriculum development, instructional design, and evaluation (Jin et al., 2019) and it has become the core concept of contemporary basic education curriculum reform (Hu, 2022). Therefore, learning progression serves as a direct guide
for process evaluation and a core tool for implementing core literacy evaluation requirements.

At the same time, research on learning progression is mainly focused on the field of STEM education, with limited attention given to foreign language education (Harris et al., 2022). A search of SSCI papers, as well as the Wanfang and China National Knowledge Infrastructure (CNKI) databases indicates that there is limited research on learning progression in Chinese language education, and no direct research specifically related to the learning progression of essay writing. The research on learning progression in English writing is also limited to persuasive writing (Deane & Yi, 2014). Therefore, it is crucial to develop and research the learning progression of Chinese essay writing, as it is an urgent task in the context of core literacy-oriented essay evaluation.

In summary, a three-strata theory for the evaluation of the core literacy can be summarized as follows: at the macro-level, the concepts and theories on the core literacies are the guiding principles; at the meso-level, academic quality standards needs to be articulated for core literacies, which can be then translated into the micro-level (or the domain/subject level) rules for core literacies’ multi-stage development through students’ whole school learning period( K to 12), which, in turn, facilitate the construction of a specific tool called “learning progression”. Based on this foundational framework, four types of evaluations are conducted, with particular emphasis on process evaluation. Therefore, the evaluation of essay writing from the perspective of core literacy includes the following specific requirements:

First, it should be guided by the idea of “core literacy” and emphasize the evaluation of advanced complex abilities.

Second, it should be based on academic quality standards and emphasize the holistic development and stage-based patterns of key literacies.

Third, it should place a strong emphasis on deep integration with new technologies.

Although automated essay evaluation technology has undergone four generations of development, the main aspects of evaluation have not seen significant changes. It mainly focuses on two dimensions: linguistic features and semantic content. Linguistic features encompass shallow language characteristics such as word frequency and sentence length, as well as deeper aspects such as grammar, syntax, and fluency. The semantic dimension primarily deals with aspects of essay content and logical structure.

It can be observed that the primary focus of automated essay evaluation is on micro-level linguistic and semantic features, which are important components of process evaluation. However, it does not directly address other requirements of intelligent evaluation for core competency outlined in the curriculum, such as advanced writing skills. Research has shown that timely feedback on essays promotes the development of intelligent writing evaluation systems (Shermis & Burstein, 2013). However, linguistic features, particularly spelling and grammar instruction, do not significantly impact the quality of writing (Graham & Perin, 2007). The critical variables are the advanced thinking abilities demonstrated in different genres of writing (Deane & Yi, 2014; Song et al., 2020). Studies by Deane and Yi (2014) on the advancement of argumentative writing align more closely with the core competency-oriented writing evaluation advocated in Chinese curriculum frameworks, but there is still a gap in the field of the Chinese language. Compared to the requirements of core competency evaluation, automated essay evaluation naturally integrates artificial intelligence technology with educational assessment and is an ideal choice for conducting intelligent evaluation based on core competencies. However, there are still several areas in which automated essay evaluation needs improvement:

First, it does not pay enough attention to the evaluation of advanced complex skills, such as organizing essay structure and logical reasoning.

Second, due to the lack of “learning progression” in essay writing, the expression of academic quality standards is insufficient, specifically manifested as:

1. Insufficient depiction of the phased development characteristics of core literacy.

2. Lack of process evaluation for the “macro” overall development laws of core literacy.

Third, there is the possibility of establishing an intelligent teaching, learning, and evaluation system based on
a unified technical framework.

In the early stages of the project, the project team cautiously focused on the automation evaluation of Chinese essays, and then based on this, appropriately considered some auxiliary functions for teachers and students, such as automated comment generation, which Chinese language teachers greatly need. Overall, due to the level of artificial intelligence technology, the project did not dare to intervene too deeply in classroom teaching or students’ independent learning, as it would require the introduction of many different artificial intelligence technologies, making the project uncontrollable and exceeding our resource capacity. However, the preliminary exploration of automated comment generation using GPT technology has solidified our determination to build a comprehensive intelligent teaching, learning, and evaluation system. In the subsequent research stage, the process of generating essay comments using GPT4 not only completely solved the problem of automated field generation but also demonstrated the potential of intelligent essay-guided instruction (as long as we design prompts that meet the needs of Chinese students in the system).

References

argumentative student essays. 29th International Joint Conference on Artificial Intelligence.


