Front. Digit. Educ., 2026, 3(2): 12
https://doi.org/10.1007/s44366-026-0086-y

RESEARCH ARTICLE

Evaluating the Efficacy of a Multifaceted Prompt for Use with LLMs to Evaluate Course Project Reports

Qingyang Sun[a,b], **Jialu Zhang**[a,b], **Peng Sheng**[a,b], **Qianyi Wang**[a,b], **Tianrui Wang**[a,b], **Heng Li**[a,b], **Hanshu Zhan**[c,d], **Xiaoqing Zhang**[a,b], **Jiang Liu**[a,b,e,f]

[a] Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China

[b] Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, Shenzhen 518055, China

[c] College of Educational Technology, Northwest Normal University, Lanzhou 730070, China

[d] Teaching Affairs Office, Southern University of Science and Technology, Shenzhen 518055, China

[e] School of Computer Science, University of Nottingham Ningbo China, Ningbo 315100, China

[f] School of Ophthalmology and Optometry, Wenzhou Medical University, Wenzhou 325035, China

Abstract The course project report (CPR) is a crucial component for assessing students' learning outcomes from courses they are studying. It assesses practical skills, academic writing, and logical thinking. In recent times, researchers have increasingly leveraged large language models (LLMs) to promote automated essay scoring (AES) in the education intelligence field due to its strong generalization and reasoning abilities. However, the existing LLM-based AES method design is based solely on writing proficiency and inevitably ignores the importance of assessment of cognitive engagement and practical competencies in CPRs. Additionally, CPR writing is a reflective process that includes knowledge-inquiry and cognition through critical thinking (CT), which have rarely been explored in the design of prompts for specific LLMs. To tackle this issue, we propose a novel, guided generative AI (GenAI) prompting framework for automated CPR assessment. It is created by integrating the Paul-Elder critical thinking concept into prompt design to enhance domain-specific knowledge transfer and the analytical capabilities of GenAI LLMs. Rather than focusing solely on language structure or writing skills, our approach emphasizes critical thinking evaluation using the Paul-Elder CT framework. Specifically, our framework—PEG-Prompt—evaluates CPR across six dimensions—structure, logic, coherence, originality, citation, and knowledge proficiency—to evaluate CPRs comprehensively from the aspects of practical competencies, analytical reasoning, and writing skills. To further enhance the CPR assessment performance of PEG-Prompt, we combine PEG-Prompt with extracted key content from reports and representative examples of few-shot scoring. Experimental results demonstrate that PEG-Prompt significantly improves the correlation between LLM-generated scores and human scores. The enhanced framework may enable students to receive helpful feedback and summaries of their CPR results through GenAI once it has been calibrated with human evaluators.

Keywords course project report assessment, Paul-Elder critical thinking concept, large language models, education intelligence

1 Introduction

The use of AI has become increasingly prevalent in various fields, including education intelligence (EI) (Chen et al., 2024; Dong et al., 2025; Wen et al., 2024). An emerging research topic, EI aims to employ AI techniques to understand the inherent characteristics of education to better assess and improve students' learning and teachers' efficacy (Liu et al., 2025). The field

Received July 22, 2025; revised October 14, 2025; accepted January 12, 2026

Xiaoqing Zhang (✉)
E-mail: xq.zhang2@siat.ac.cn
Jiang Liu (✉)
E-mail: liuj@sustech.edu.cn

addresses multiple dimensions of education: homework, course project reports (CPRs), course learning, classroom instruction, and academic discussions. CPR stands out as a particularly vital intermediate component in higher education; it bridges the gap between learning and teaching through continuous feedback and interactions. In the Chinese higher education context, a CPR generally refers to a written report that documents the process, outcomes, and reflections related to a course-related project. It integrates practical implementation with academic writing, serving as both an evaluation of students' mastery of course knowledge and as a training exercise for research and communication skills. These written course reports encompass various competencies—from course content comprehension and academic writing standards to practical skills, innovation, and logical reasoning (Friesen & Scott, 2013; Lee, 2014; Pedaste et al., 2015; Zhang et al., 2021). It is well-acknowledged that CPR plays an essential role in evaluating students' learning outcomes, which not only reflect students' proficiency in learning the course content but also reveal their academic writing skills and logical thinking abilities. Teachers often provide corresponding scores and comments on the CPRs, which is time-consuming, labor-intensive, and subjective. Students then reflect on and address their weaknesses according to the teacher's feedback, thereby improving their mastery of course-related knowledge as well as enhancing their academic writing and logical thinking skills. Current CPR-related studies have predominantly examined implementation practices and their quantifiable results through conventional evaluation approaches—questionnaires, feedback analysis, and performance metrics (Justice et al., 2007; Mieg, 2019; Nabhan, 2017; Waked et al., 2024). However, these investigations have failed to address something fundamental: the intrinsic quality of CPRs assessment. This becomes particularly significant given that no prior research has explored assessment of CPRs from the EI perspective.

Recently, there has been promising progress with using large language models (LLMs) in automated essay scoring (AES) due to their robust reasoning and content generation capabilities. For example, researchers have used ChatGPT to analyze students' essays, demonstrating the potential of using LLMs in AES (Liu et al., 2024b; Mizumoto & Eguchi, 2023; Sun et al., 2025). Han et al. (2023) developed the RECIPE platform, which integrates ChatGPT to assess foreign language learners' writing; it received positive feedback from most students. Yancey et al. (2023) used various prompt engineering strategies to score short L2 essays on the common European framework of reference scale, showing the potential of this approach in enhancing LLM performance on downstream tasks. Stahl et al. (2024) also investigated the effects of different prompting strategies on automated essay scoring and feedback generation. Although these LLM-based AES methods have achieved good results, employing them in CPR assessment has not been explored previously.

It is worth noting that writing a CPR is a reflective process that includes knowledge inquiry and cognition stages, which can be viewed as another form of critical thinking (CT). This is mainly because CT involves reasoned analysis of information, systematic evaluation of arguments, and formulation of well-substantiated judgments (Ennis, 1993; Facione, 1990; Halpern, 2013). Moreover, the Paul-Elder CT framework deconstructs the classical CT into tripartite interactive components: intellectual standards, elements of reasoning, and intellectual traits. They work together as follows: intellectual standards are applied to elements of reasoning, thereby fostering the development of intellectual traits, which collectively contribute to the growth of higher-order thinking (Paul & Elder, 2019). Drawing on this framework, we propose integrating Paul-Elder CT into prompt design to better evaluate students' cognitive engagement in automated CPR assessment. This integration enables LLMs to extract specific domain knowledge and reasoning abilities, facilitating more automated—yet precise—CPR assessment.

Motivated by the above systematic analysis, we propose a novel Paul-Elder CT-guided prompting framework—PEG-Prompt—based on Paul-Elder CT and writing proficiency, which aims to efficiently leverage the domain-specific knowledge and reasoning ability of LLMs for automated CPR assessment. Specifically, this paper presents the design of the PEG-Prompt based on six different dimensions that draw on the Paul-Elder CT concept—these dimensions are structure, logic, coherence, originality, citation, and knowledge proficiency. Furthermore, we extract key content from reports and select representative example of few-shot scoring, which are combined with PEG-Prompt to boost LLMs' CPR assessment performance. Figure 1 compares how inputting a base prompt into an LLM compares with putting a PEG-Prompt into an LLM to assess CPRs. It can be observed that with the guidance of the base prompt, LLMs only output the general score of a CPR. With the guidance of our PEG-Prompt, LLMs not only output the overall score of a CPR but also generate a detailed assessment of a CPR from six dimensions. This comprehensive evaluation can assist teachers in conducting effective and efficient CPR assessments while enabling students to identify their shortcomings and pathways for improvement.

The main contributions of this paper are summarized as follows:

1) Inspired by the close link between the process of writing CPRs and CT, this paper is the first to propose the novel PEG-Prompt to guide LLMs to

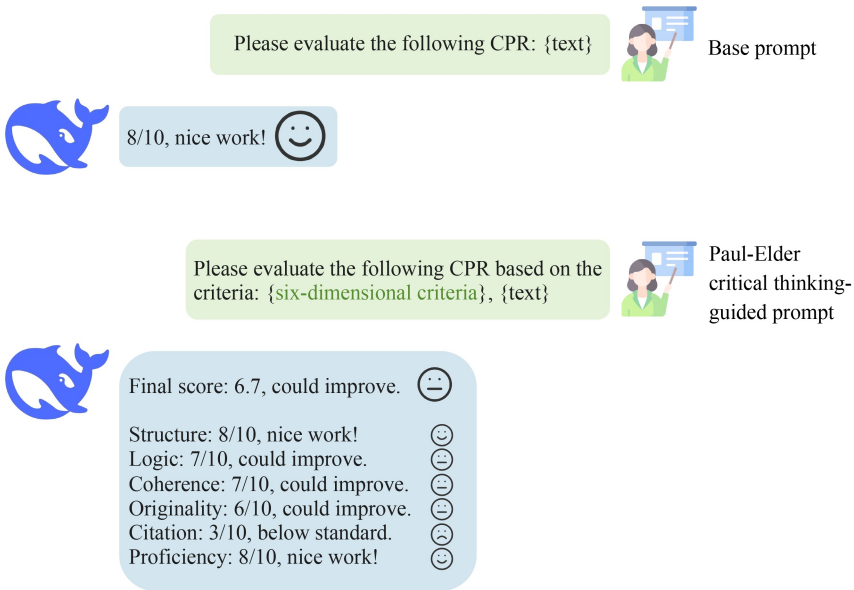

Figure 1 Comparison between assessment of a CPR using a base prompt and using PEG-Prompt. Upper: Using base prompts in an LLM exhibits a limitation—they assess CPR solely from the perspective of writing proficiency, thereby constraining evaluation accuracy; lower: In contrast, our proposed PEG-Prompt framework incorporates critical thinking, guiding the LLM in generating comprehensive assessments across six distinct dimensions. CPR: course project report, LLM: large language model.

produce a comprehensive CPR assessment by incorporating the Paul-Elder CT concept into prompt design.

2) We further integrate key report content and representative few-shot scoring examples into PEG-Prompt to enhance the performance of LLMs in assessing CPRs.

3) We construct a CPR data set—PEG-CPR—to advance the development of the CPR assessment field. Experimental results demonstrate that our method enhances the performance of state-of-the-art (SOTA) LLMs, aligning with our expectations.

2 Related Work

2.1 | Automated Essay Scoring

Automated essay scoring has been extensively studied as a natural language processing (NLP) task that aims to assign holistic scores to student essays based on predefined scoring criteria (Li & Ng, 2024; Ramesh & Sanampudi, 2022). Traditional AES, e.g., project essay grade (PEG) and electronic essay rater (E-rater) relied heavily on heuristic but rule-based methods (Attali & Burstein, 2006; Page, 1967) with extensive manual feature engineering. However, these methods are time-consuming and only allow subjective grading. With the advent of AI, researchers have heavily promoted the development of automated AES. For instance, while Ke et al. (2019) still employed manual feature extraction, they advanced the field by applying a support vector machine (SVM) to systematically calculate essay scores based on higher-level semantic features such as argument strength, specificity, and clarity. Cader (2020) combined data augmentation techniques with deep neural networks to improve AES performance. More recently, researchers have used LLMs to perform AES. Mizumoto and Eguchi (2023) demonstrated that LLMs are capable of performing holistic essay scoring by leveraging their extensive linguistic knowledge and commonsense reasoning abilities guided by complex prompt instructions. Subsequently, Mansour et al. (2024) investigated how prompt engineering impacts AES performance, revealing that while LLMs underperform in comparison to SOTA methods in scoring accuracy, they show promise in generating constructive feedback to support improvement of students' writing improvement. Xiao et al. (2024) systematically explored prompt engineering strategies for AES, revealing that carefully tailored prompts significantly boost the scoring performance and feedback quality of GPT models, further supporting human evaluators in educational contexts. Unfortunately, these methods primarily focus on writing skill evaluation, which has limitations in that it does not investigate the underlying thinking and insights of students in writing their CPRs.

2.2 | Course Project Report Assessment

Course project report refers to a report that integrates practical processes with written expression to present research findings and reflections (Justice et al., 2007).

Mieg (2019) conducted a comprehensive investigation into the integration of research activities into undergraduate curricula in German higher education; it encompassed pedagogical approaches, disciplinary case studies, and assessment methods. Nabhan (2017) explored the integration of a student-centered inquiry approach and blogging as part of writing instruction, finding through questionnaires and interviews that students held a positive attitude toward this learning method. Waked et al. (2024) have shown that CPRs can improve academic performance and student engagement. Current studies have explored the implementation of CPRs in classroom instruction and have analyzed effectiveness through student feedback or academic performance improvement. However, these studies predominantly focus on pedagogical design and learning outcomes assessment. Our extensive survey found that existing studies have not employed LLMs to assess the quality of CPRs from the EI perspective.

2.3 | Critical Thinking

Critical thinking, as defined by Facione (1990), encompasses the abilities of interpretation, analysis, evaluation, inference, explanation, and self-regulation. Ennis (1993) describes CT as reasonable reflective thinking that focuses on deciding what to believe or do and highlights dispositional elements along with cognitive skills. Halpern (2013) conceptualizes CT as thinking that is purposeful, reasoned, and goal-directed. Notably, Paul and Elder's (2019) theoretical framework presents a comprehensive model comprising three fundamental components—intellectual standards, elements of reasoning, and intellectual traits—which, collectively, serve to cultivate and enhance higher-order thinking competencies. More closely related to our work—considering that CPR assessment necessitates the evaluation of reasoning processes—the Paul-Elder CT framework provides an essential methodological framework for CPR evaluation. Therefore, this paper integrates the Paul-Elder CT concept into LLM prompt design to enhance the domain-specific knowledge transfer and analytical capabilities of LLMs in CPR assessment.

3 Methodology

This section begins with a definition of automated CPR assessment, followed by a detailed introduction to our PEG-Prompt. Additionally, we introduce prompt engineering optimization through extracting key content from reports and providing examples of few-shot scoring.

3.1 | Task Definition

As introduced previously, the main goal of this research is to develop an efficient prompt to allow LLMs to produce an objective yet comprehensive CPR assessment from various viewpoints. Formally, we construct a function F that maps a CPR R_{cpr} to a multidimensional assessment E_{out}:

$$F(R_{\text{cpr}}) = E_{\text{out}} = S, R, \sigma, \phi, \tag{1}$$

where $S = i_1, i_2, \ldots, i_6$ denotes the scores across six dimensions, $R = r_1, r_2, \ldots, r_6$ represents corresponding rationales, σ represents the weighted overall score, and ϕ provides comprehensive improvement suggestions.

3.2 | Paul-Elder Critical Thinking Guided Prompting

As previously noted, CPR writing involves both engaging in academic writing and showing that one can think logically. However, existing studies have not sufficiently integrated CT traits into the design of evaluation criterion. Drawing on the Paul-Elder CT concept, we have developed the Paul-Elder critical thinking guided prompting (PEG-Prompt) framework, which enables LLMs to assess CPRs by jointly considering writing proficiency and CT abilities. As illustrated in Figure 2, the Paul-Elder CT concept breaks down CT as a cognitive process comprising intellectual standards, elements of thought, and intellectual traits. Specifically, intellectual standards serve as a criterion for assessing quality of thinking, such as clarity, accuracy, and logic. Elements of thought represent the fundamental building blocks of reasoning, encompassing purpose, inference, questions, and so on. The core mechanism of the Paul-Elder CT concept is employing intellectual standards to evaluate elements of thought for developing intellectual traits—the dispositions and attitudes manifested as a person's thinking processes, such as intellectual integrity, intellectual perseverance, and confidence in reason. Similarly, the CPR writing process involves the expression and presentation of elements of thought. Therefore, drawing on the evaluative principles of intellectual standards, we summarize three core dimensions for assessing CT levels in CPR writing: originality, logic, and knowledge proficiency. Additionally, in consideration of the academic standards of CPRs, students' writing competence and academic writing conventions are also integrated into the evaluation framework. Consequently, coherence of language, report structure, and citation compliance are also adopted in our assessment criteria.

In light of the above analysis, we propose six novel evaluation criteria to embed into the design of PEG-Prompt, as shown in Figure 3. Specifically, originality examines students' unique insights or novel

solutions based on existing technologies; logic evaluates whether the CPR demonstrates coherent reasoning in terms of problem analysis and solution selection; and knowledge proficiency measures the depth of students' theoretical understanding and practical applications of the knowledge. These three dimensions are designed to assess CT abilities; that is, students' higher-order cognitive skills as demonstrated in the CPR. Meanwhile, coherence examines the accuracy of terminology used and fluency of expression; structure evaluates whether the CPR is organized clearly; and citation assesses whether the students have correctly cited relevant theories and technical literature to ensure academic rigor. These three dimensions constitute the core of traditional writing assessment. By integrating CT dimensions with traditional writing assessment metrics, the evaluation comprehensively analyzes both logical thinking and academic expression in CPRs.

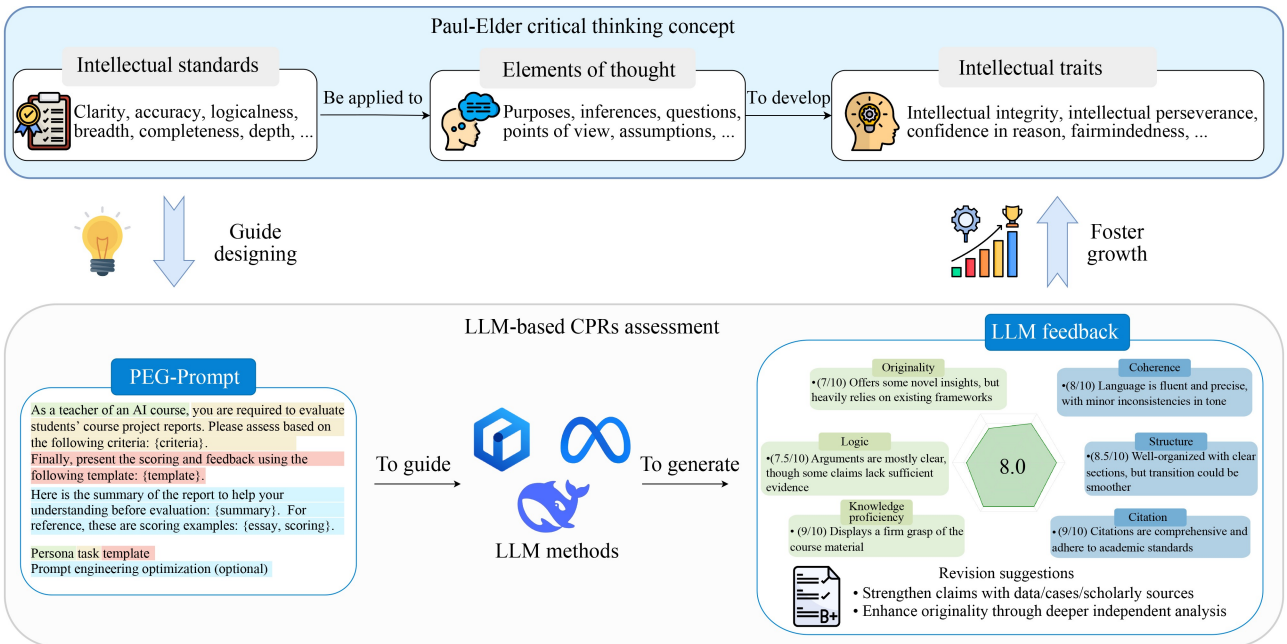

Figure 2 General framework of PEG-Prompt—which is inspired by the Paul-Elder critical thinking framework—incorporates six dimensions. This framework guides LLMs in evaluating CPR with respect to practical competencies, analytical reasoning, and writing skills while generating targeted feedback. This feedback facilitates students' ability to reflect on their results and improve in their weaker areas, thereby contributing to the cultivation of higher-order intellectual traits.

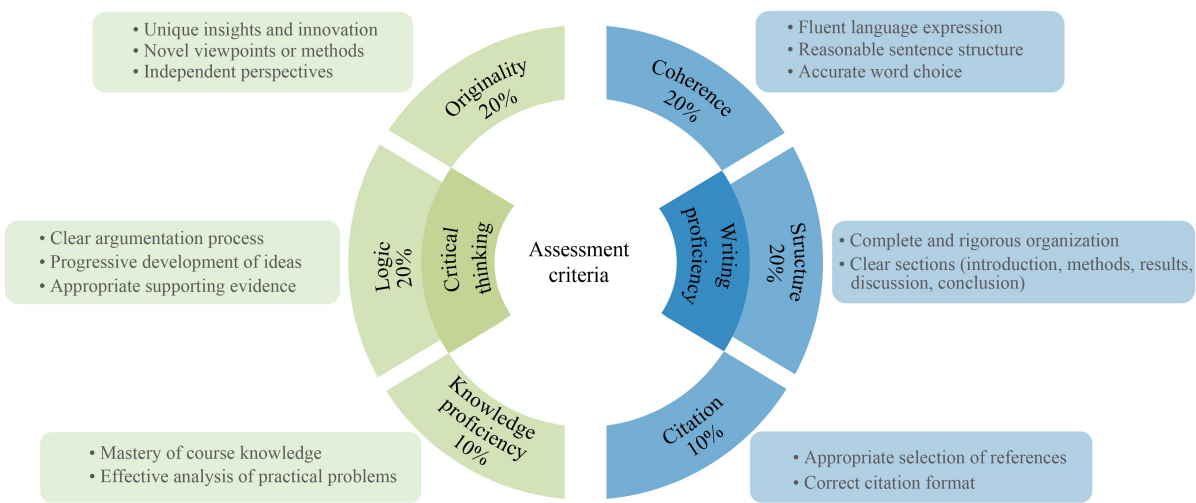

Figure 3 Six-dimensional evaluation criteria are introduced in the PEG-Prompt, which is informed by the Paul-Elder critical thinking concept and writing proficiency. The aim is to assess the quality of CPRs comprehensively and objectively.

As shown in Figure 2, building upon Paul-Elder CT concept-inspired assessment criteria, we have designed our PEG-Prompt following established prompt design paradigms (Amatriain, 2024; Cao et al., 2025). The prompt design incorporates persona-setting (White et al., 2023), which positions the LLM as a university teacher in an AI course; the aim is to implicitly guide the LLM to focus on domain-specific knowledge while employing professional and pedagogical language. Then, the prompt presents task instructions that require the LLM to evaluate the CPR across six dimensions, assigning a score from 0 to 10 for each and providing corresponding rationales. Finally, predefined templates guide the model in generating structured outputs. Our PEG-Prompt guides the LLM to generate accurate and formatted feedback, including the overall score, the scores and justifications for each of the six dimensions, and general suggestions for revision. Through comprehensive assessment feedback, students can reflect on their deficiencies and implement targeted improvements, thereby enhancing their writing proficiency while fostering the growth of intellectual traits essential for CT. Though the feedback can support formative learning by guiding revisions and self-reflection, it may also be adapted for summative assessment when integrated into a formal course evaluation.

3.3 | Prompt Engineering Optimization

To enhance the CPR assessment performance of our PEG-Prompt, we introduced two prompt engineering optimization strategies: 1) extraction of key content from reports for providing informative report content to PEG-Prompt with the help of text-summarization techniques; 2) curation of few-shot scoring examples for establishing standard CPR assessment patterns in the PEG-Prompt by carefully curating representative few-shot scoring examples.

3.3.1 Extraction of Key Content From Reports

Extracting key content from reports is helpful in improving the CPR assessment performance of PEG-Prompt with LLMs. Therefore, to extract key content from a CPR, we constructed a multi-stage process with text summarization techniques, as illustrated in Figure 4.

First, we perform *text chunking* to split the CPR into overlapping segments, ensuring semantic continuity across adjacent chunks.

Second, each chunk is encoded into a high-dimensional vector with a pre-trained embedding model f_{embed}. Taking a set of chunked texts $\{c_1, c_2, \ldots, c_n\}$, we compute their corresponding embeddings as:

$$\boldsymbol{v}_i = f_{\text{embed}}(c_i), \quad \forall i \in [1, n], \tag{2}$$

These vectors are stored in a vector database $\boldsymbol{v}$ to support similarity-based retrieval.

Third, when an extraction query is initiated, maximum marginal relevance (MMR) is applied to retrieve top-k relevant yet diverse text segments. MMR balances relevance and diversity by iteratively selecting the next chunk c^* that maximizes MMR objective function, defined as:

$$c^* = \arg\max_{c_i \in C \setminus S} \left[\lambda \cdot sim(\boldsymbol{q}, \boldsymbol{v}_i) - (1-\lambda) \cdot \max_{c_j \in S} sim(\boldsymbol{v}_i, \boldsymbol{v}_j) \right], \tag{3}$$

where C is the candidate set, S is the current set of selected chunks, $\boldsymbol{q}$ is the query vector, and $\lambda \in [0, 1]$ controls the trade-off between relevance and diversity.

Finally, the retrieved segments are synthesized into concise, extracted key report content that captures the core information about the CPR. Then, the extracted content is subsequently integrated, as a structured component, into the PEG-Prompt, facilitating more context-aware and criteria-aligned assessment with LLMs.

3.3.2 Curation of Few-Shot Scoring Examples

Few-shot prompting is a widely used technique to improve the general performance of LLMs working on specific tasks. It involves inserting a few representative examples into the prompt. Motivated by the success of curated example selection in few-shot prompting design, we carefully selected representative CPR scoring examples. These examples were integrated into

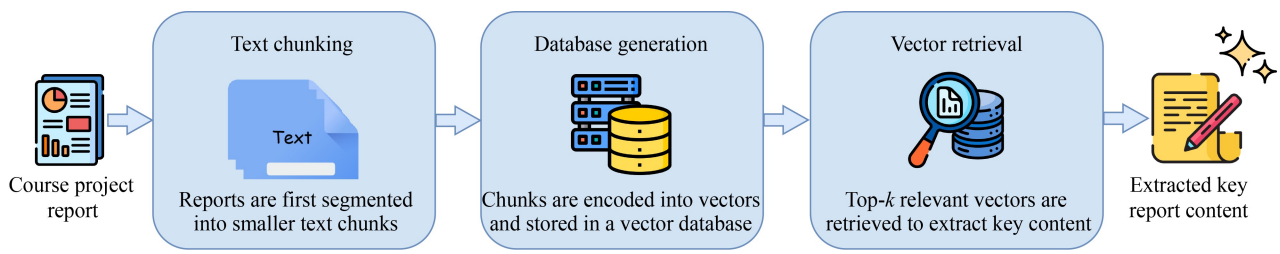

Figure 4 Multi-stage process to extract key content from a CPR.

PEG-Prompt to enhance assessment performance by providing LLMs with specific scoring and reasoning processes. Note that these representative scoring examples encompassed multidimensional ratings with detailed justifications, comprehensive overall scores, and revision suggestions in an instructional format. The structured design allowed us to fine-tune inputs for our PEG-Prompt; this enables LLMs to produce accurate and high-quality CPR assessment outputs that better align with assessment by human teachers.

4 Experiments

4.1 | Dataset

This study introduces the PEG-CPR dataset, comprising CPRs collected from the Introduction to Artificial Intelligence course at the Southern University of Science and Technology during the 2022–2023 academic year. Dataset collection was conducted in accordance with institutional ethical guidelines for educational research. Due to the retrospective design and fully anonymized processing—where all personally identifiable information, including names, student IDs, and email addresses was removed—the university's academic ethics committee granted an exemption from individual consent requirements. The course attracted students from diverse academic backgrounds, including computer science, medicine, finance, and other disciplines, reflecting the interdisciplinary nature of AI applications. The PEG-CPR dataset comprises 110 high-quality CPRs filtered from an initial 117 submissions. Reports exhibiting severe formatting issues or that had insufficient content were excluded from the data set to ensure reliability of evaluation. Project topics were derived from real-world AI application scenarios, such as the development of intelligent cataract screening systems and AI-powered Go games, allowing students to explore practical implementations of AI technologies across various domains. Students were allowed to write their project reports in either Chinese or English, based on their preference and expertise.

Each report was assessed by experienced course instructors using the six-dimensional assessment framework—they assessed for structure, logic, coherence, originality, citation, and knowledge proficiency. To ensure methodological rigor, all reports were independently scored by two instructors. Each instructor was responsible for approximately half of the total reports, and a subset of 20 reports was double-scored to check consistency. Inter-rater reliability was measured using Cohen's kappa, which yielded a coefficient of 0.82, indicating strong agreement. Any discrepancies in the double-scored subset were resolved through discussion. This procedure ensured both the reliability and validity of the human evaluation process. The evaluations were conducted in Chinese to maintain consistency and standardization across all assessments. A final, holistic score was computed for each report as a weighted aggregate across the dimension-level scores. As shown in Figure 5, the overall CPR score distribution approximates a normal distribution, with most scores falling within the range of 7.6–9.0. This concentration around the middle range aligns well with typical student performance patterns, where the majority demonstrate intermediate-level competencies. Moreover, the distribution demonstrates clear differentiation across the scoring spectrum, showing a smooth gradient from lower to higher performance levels without clustering at extreme values, which further confirms the discriminative power of our assessment framework. This balanced distribution pattern indicates that the evaluation criteria effectively capture performance variations while maintaining scoring rationality. Detailed statistics related to the PEG-CPR are presented in Table 1.

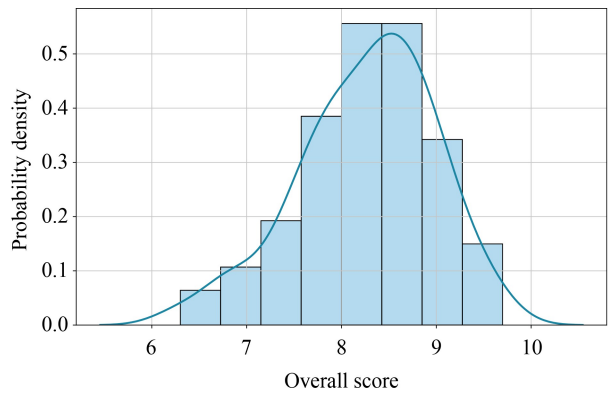

Figure 5 Histogram of overall CPR assessment scores. The distribution follows an approximately normal curve.

Table 1 Mean scores across six dimensions and overall score in the PEG-CPR dataset

Dimension	Mean score
Structure	8.44
Logic	8.25
Coherence	8.23
Originality	8.20
Citation	7.49
Knowledge proficiency	8.91
Overall	8.26

4.2 | Evaluation Metrics

The performance of the proposed method was evaluated using three commonly used metrics: arithmetic mean (Mean), mean absolute error (MAE), and mean

squared error (MSE). Specifically, mean captures the central tendency of the scores. MAE measures the average absolute deviation between predicted scores and human-assigned scores. MSE further emphasizes larger discrepancies by squaring the errors, providing a more sensitive evaluation of scoring consistency. These metrics are defined as follows.

$$\text{Mean} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{4}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |x_i - \widehat{x}_i|, \tag{5}$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \widehat{x}_i)^2, \tag{6}$$

where N is the total number of evaluated reports and x_i and $\widehat{x}_i$ denote the predicted and human-assigned scores for the i-th report, respectively. All metrics—mean, MAE, and MSE—were computed for the overall score to comprehensively evaluate model performance. For the six individual assessment dimensions, we reported the mean scores and examined their closeness to human teacher scores. In this study, better performance is defined as closer alignment between LLM-predicted and human teacher scores, which is quantitatively reflected in lower MAE and MSE for the overall score and qualitatively supported by dimension-level consistency and report-level improvements. This definition provides the basis for the subsequent performance comparisons reported in Section 5.

4.3 | Baselines

To evaluate the effectiveness of the proposed PEG-Prompt, we conducted comparative experiments using four representative LLMs, all of which are proprietary, closed-source systems accessed via official APIs. The selected LLMs were:

1) **ERNIE Speed** (Sun et al., 2021): Developed by Baidu, this is a multilingual LLM with strong capabilities in Chinese-language understanding and long-context processing. It is designed for efficient performance in general-purpose, open-domain applications rather than specialized or domain-specific tasks. For simplicity, we refer to it as ERNIE Bot throughout this paper.

2) **DeepSeek-V3** (Liu et al., 2024a): A Chinese high-performance mixture of experts model. Although it is optimized for general-domain tasks, its advanced capabilities in reasoning, content organization, and factual consistency enable it to generate more reliable and coherent responses in complex scenarios such as automated CPR assessment. We refer to this model as DeepSeek through the rest of this paper.

3) **Gemini 1.5 Flash** (Gemini Team, Google, 2024): Google's multi-modal foundation model; it excels at understanding text, image, and code. It exhibits robust linguistic comprehension and logical inference capabilities, making it particularly suitable for structured evaluation scenarios. Throughout this paper, we call this model Gemini.

4) **Llama Guard 3 8B** (Grattafiori et al., 2024): An LLM developed by Meta, designed with an emphasis on parameter efficiency and computational optimization. While originally trained primarily on English corpora, Llama exhibits strong transfer learning ability and can adapt effectively to multilingual contexts through fine-tuning. We use Llama as the abbreviated reference for this model going forward.

4.4 | Implementation Details

All experiments were conducted using proprietary LLMs accessed via APIs. Specifically, the prompts and reports were provided as text input to the API in the same way as ordinary user queries, and the models returned the corresponding outputs in JSON or text format, which do not require any fine-tuning or system-level modification. To ensure experimental stability and reproducibility, the temperature parameter was fixed at 0 across all experimental settings. For the extraction of key content from reports, input texts were segmented into 1000-character chunks with a 200-character overlap to maintain contextual coherence. Embeddings were generated using the all-mpnet-base-v2 model and stored in a ChromaDB vector database. Chunk retrieval was configured to return the top 8 relevant segments from a candidate pool of 20, with the trade-off parameter λ set to 0.6. The retrieved chunks were subsequently passed to the generation model, with temperature set to 0.3 and top-p sampling set to 0.9 to balance generation diversity and output coherence. These hyperparameters were adjusted via the API call, thereby reducing output randomness and enhancing reproducibility of the generation results. For the curation of few-shot scoring examples, the number of report-score pairs used to construct the prompt was set to 5. These examples were excluded from the evaluation phase to eliminate any potential evaluation bias resulting from example reuse.

For comparative analysis, we implemented a base prompt that served as our baseline. The base prompt followed a similar structure to PEG-Prompt in that it set up the LLM as a university teacher in an AI course, but it differed significantly in its task instructions. Specifically, the base prompt requested that the LLM directly evaluate the input report and provide only an overall holistic score without multidimensional assessment or detailed rationales. Unlike PEG-Prompt's

six-dimensional framework previously defined in this paper, the base prompt did not incorporate the Paul-Elder critical thinking concept or require dimension-specific evaluations. This design choice allowed us to isolate the contribution of our proposed multidimensional assessment framework and demonstrate how PEG-Prompt enhances the capability of LLMs in providing comprehensive CPR evaluation. For reproducibility, the scoring rubrics adopted in this study were clearly defined by the six assessment dimensions and their criteria, as illustrated in Figure 3 and described in detail in Subsection 3.2. They provided explicit guidance on how both human instructors and the LLMs evaluate the course project reports.

5 Results and Analysis

5.1 | Effectiveness Validation

5.1.1 Effectiveness Analysis of PEG-Prompt

Table 2 presents a comprehensive comparison of various LLM-based methods using both the base prompt and the proposed PEG-Prompt approach. Results show that PEG-Prompt consistently boosts LLM performance compared to the base prompt across multiple evaluation metrics. All models improved with PEG-Prompt, with DeepSeek achieving an average overall score of 8.13, closely aligning with the human teacher evaluation of 8.26. Performance gains include reduced error rates, with MAE decreasing from 0.73 to 0.57 and MSE reducing from 1.31 to 0.73. Gemini delivered the most significant improvement, narrowing the expert score gap by 0.61 points and reducing MAE from 1.94 to 1.20 and MSE from 5.85 to 2.70. Low standard deviations across PEG-Prompt results indicate high consistency and reliability in scoring accuracy. These findings collectively demonstrate that PEG-Prompt enhances both scoring accuracy and prediction stability across diverse model architectures.

Beyond overall score improvements, PEG-Prompt's most significant contribution is enabling structured multidimensional assessments, as opposed to the base prompt, which yield only holistic scores. As shown in Table 2, PEG-Prompt provides assessments across the six critical dimensions previously outlined in this paper. These dimension-level evaluations highlight the varying strengths of each model across our multidimensional assessment criteria. For example, while DeepSeek leads in most dimensions, Gemini performs marginally better in citation.

5.1.2 Analysis of the Effectiveness of Prompt Engineering Optimization

To evaluate the effectiveness of the prompt engineering optimization, we designed three experimental configurations: 1) "Key content", incorporating algorithmically extracted report summaries, as described in Subsection 3.3.1, to provide focused contextual guidance; 2) "Scoring examples", where we implement few-shot learning—here, the LLM learns from a small set of teacher-annotated reports to assess remaining submissions, as described in Subsection 3.3.2; and 3) "Both", combining both strategies. Notably, few-shot examples are excluded from the test set to ensure genuine generalization rather than memorization. Table 3 evaluates the impact of different prompt engineering optimization strategies applied to LLM-based methods using PEG-Prompt, focusing on ERNIE Bot and DeepSeek due to their superior baseline performance, shown in Table 2. All optimization strategies consistently improved performance across both models. Integrating extracted key content from reports ("Key content") delivered enhancements in evaluation accuracy for both ERNIE Bot and DeepSeek, with notably reduced MAE and

Table 2 Performance comparison of LLM-based methods using the base prompt and the proposed PEG-Prompt

Method	Overall score			Mean score					
	Mean	MAE↓	MSE↓	Structure	Logic	Coherence	Originality	Citation	Knowledge proficiency
Human teacher	8.26 ± 0.72	–	–	8.44	8.25	8.23	8.20	7.49	8.91
ERNIE Bot_{Base}	7.76 ± 0.87	0.85	1.05	–	–	–	–	–	–
ERNIE Bot_{PEG}	7.70 ± 0.93	**0.68**	**1.01**	8.15	7.76	7.64	7.37	8.03	8.03
$DeepSeek_{Base}$	8.09 ± 1.18	0.73	1.31	–	–	–	–	–	–
$DeepSeek_{PEG}$	8.13 ± 0.90	**0.57**	**0.73**	**8.52**	**8.07**	**7.82**	**8.03**	7.89	**8.65**
$Gemini_{Base}$	6.46 ± 1.45	1.94	5.85	–	–	–	–	–	–
$Gemini_{PEG}$	7.07 ± 1.06	**1.20**	**2.70**	7.69	6.65	6.92	5.55	**7.81**	7.18
$Llama_{Base}$	7.66 ± 1.11	0.91	1.72	–	–	–	–	–	–
$Llama_{PEG}$	7.84 ± 0.93	**0.70**	**0.90**	8.02	7.80	7.75	7.24	8.31	7.86

Notes. Subscript denotes prompt types: "Base" for base prompt, "PEG" for PEG-Prompt. Human teacher evaluations are included for reference. The best results are in bold.

Table 3 Performance comparison of LLM-based methods with PEG-Prompt using three different prompt engineering optimization strategies

Method	Optimization strategy	Overall score			Mean score					
		Mean	MAE↓	MSE↓	Structure	Logic	Coherence	Originality	Citation	Knowledge proficiency
Human teacher	–	8.26 ± 0.72	–	–	8.44	8.25	8.23	8.20	7.49	8.91
ERNIE Bot$_{PEG}$	–	7.70 ± 0.93	0.68	1.01	8.15	7.76	7.64	7.37	8.03	8.03
	+ Key content	7.82 ± 0.49	0.63	0.72	8.20	7.93	7.65	7.35	**7.74**	8.36
	+ Scoring examples	8.41 ± 0.89	0.67	0.76	8.60	8.50	**8.32**	8.66	8.00	**8.74**
	+ Both	8.31 ± 0.60	**0.46**	**0.29**	**8.60**	**8.39**	8.36	**8.11**	7.99	8.72
DeepSeek$_{PEG}$	–	8.13 ± 0.90	0.57	0.73	8.52	8.07	7.82	**8.03**	7.89	8.65
	+ Key content	8.18 ± 0.48	0.57	0.55	8.66	8.05	8.16	7.65	7.30	8.77
	+ Scoring examples	8.59 ± 0.67	0.51	0.37	8.81	8.57	8.43	8.44	8.44	8.79
	+ Both	8.27 ± 0.62	**0.41**	**0.23**	**8.51**	**8.31**	**8.31**	8.00	**7.28**	**8.83**

Notes. "Key content" for incorporating extracted key content from reports, "Scoring examples" for providing examples of few-shot CPR scoring, and "Both" for combining both strategies. Human teacher annotations are provided for reference. The best results are highlighted in bold.

MSE. This improvement highlights the effectiveness of providing models with concise extracted contextual information. Similarly, the curation of few-shot scoring examples ("Scoring examples") significantly improves alignment with human-teacher assessments. Notably, this strategy demonstrates that the PEG-enhanced framework can achieve enhanced performance without requiring teachers to grade every submission. The effectiveness of this optimization strategy is particularly evident in DeepSeek. When compared to the variant that utilizes only the proposed PEG-Prompt, the MAE decreases from 0.57 to 0.51, while the MSE is reduced by 0.26, reaching 0.37. These results demonstrate the value of few-shot scoring examples in clarifying evaluation standards and supporting structured reasoning. The combination of both strategies ("Both") yielded the most compelling results. Specifically, ERNIE Bot achieved an overall score of 8.27, closely matching the human teacher benchmark of 8.26, with an MAE of 0.41 and MSE of 0.23. DeepSeek demonstrated a similarly strong performance, achieving an overall score of 8.31 with comparably low variance.

Multidimensional scores further validate the effectiveness of prompt engineering optimization in achieving human-like evaluation capabilities. As shown in Table 3, the combined optimization strategy consistently produces scores that closely align with human teacher evaluations across all six dimensions. For DeepSeek, the combined approach yields structure at 8.51 (vs. 8.44), logic at 8.31 (vs. 8.25), coherence at 8.31 (vs. 8.23), citation at 7.28 (vs. 7.49) and proficiency at 8.83 (vs. 8.91). ERNIE Bot also shows strong alignment, particularly in originality (8.11 vs. 8.20). These consistent improvements in both overall and dimension-level performance highlight the effectiveness of prompt optimization in enhancing alignment with expert evaluations and enabling more fine-grained and reliable assessments of CPR.

5.2 | Visualization Analysis and Statistical Significance Analysis

5.2.1 Visualization Analysis

To visually assess the effectiveness of the proposed PEG-Prompt enhanced through prompt engineering optimization, we measured the Euclidean distances between the teacher's evaluations and the LLM-generated scores using both the base prompt and the PEG-Prompt. The Euclidean distance quantifies the overall scoring discrepancy between human raters and the LLM on a given essay, with smaller values indicating higher agreement and thus better alignment with human judgment. Figure 6 displays the Euclidean distance histogram; kernel density estimation (KDE) curves; and the corresponding cumulative distribution function (CDF) curves, which are obtained by integrating the KDE curves.

A leftward shift in the histogram and a KDE peak closer to the origin indicate an increased similarity between the LLM-predicted scores and the human-assigned ratings. Correspondingly, a steeper CDF curve reflects a higher cumulative density in the lower distance range, signifying closer alignment with human evaluations. As illustrated in Figure 6(a), the histogram of ERNIE Bot instructed by PEG-Prompt exhibits a notable leftward shift compared to its base prompt counterpart, indicating the superiority of PEG-Prompt. Similarly, in Figure 6(b), DeepSeek with PEG-Prompt achieves a KDE peak around 0.35, substantially lower than the 0.6 observed with the base prompt. Furthermore, the CDF curves for PEG-Prompt-enhanced models exhibit consistently steeper slopes, demonstrating that PEG-Prompt facilitates better alignment between LLM outputs and human scores.

5.2.2 Statistical Significance Analysis

To quantify the statistical significance of performance improvements enabled by our PEG-Prompt after applying prompt engineering optimization, we conducted independent two-sample *t*-tests comparing the score distributions of LLMs using the optimized PEG-Prompt (combining both key content extraction and few-shot scoring examples) with those using the base prompt. The null hypothesis assumes no significant difference between the two distributions. The resulting *t*-statistics and *p*-values are summarized in Table 4.

For ERNIE Bot, the PEG-Prompt variant yielded a *t*-statistic of 6.6588 with a *p*-value of 1.1913×10^{-9}, which is well below the standard significance threshold of 0.05. Similarly, the DeepSeek comparison resulted in a *t*-statistic of 2.2995 and a *p*-value of 0.0234, also indicating statistical significance. These results decisively reject the null hypothesis at a 5% confidence level, confirming that PEG-Prompt significantly improves the alignment between LLM-generated scores and human evaluations.

5.3 | Case Study

An in-depth comparative analysis of DeepSeek-V3's

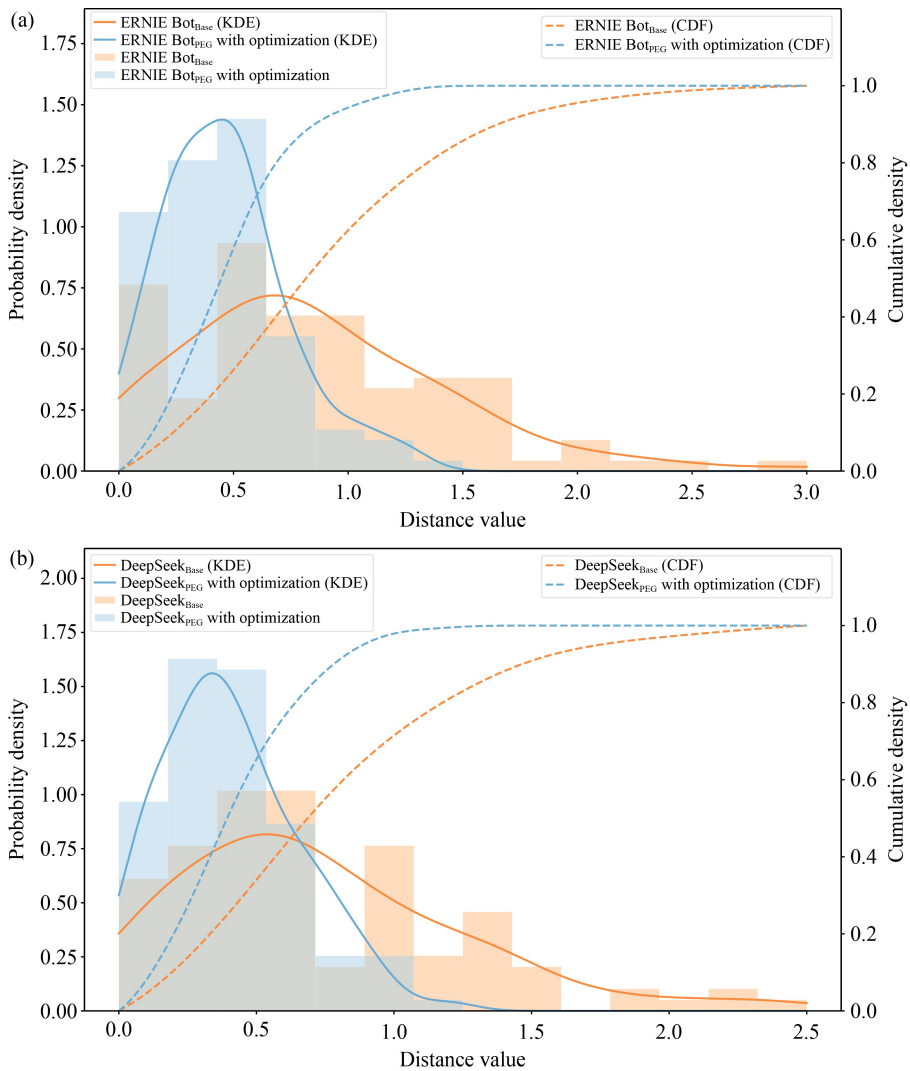

Figure 6 Euclidean distance distributions between scores assigned by human teachers and predicted by LLMs using the base prompt and PEG-Prompt. (a) ERNIE Bot; (b) DeepSeek.

Table 4 *t*-statistics and *p*-values for comparisons between PEG-Prompt and base prompt variants

Method	*t*-statistic	*p*-value
ERNIE Bot$_{PEG}$ with optimization vs. ERNIE Bot$_{Base}$	6.6588	1.1913×10^{-9}
DeepSeek$_{PEG}$ with optimization vs. DeepSeek$_{Base}$	2.2995	0.0234

evaluations under different prompting strategies is presented in Figure 7, based on the same student report. The original Chinese text was translated into English for clarity. Specifically, when employing the base prompt, the model primarily focused on surface-level writing features, resulting in limited assessment validity. Moreover, the outputs generated often deviated from the expected response format, complicating downstream interpretation and automated processing. In contrast, with the PEG-Prompt, the model evaluated both critical thinking and writing quality across the six defined dimensions. This led to improved content relevance and structure. However, certain scores and rationales still deviated from the teacher's evaluation, and the feedback remained insufficiently detailed. With further enhancement through prompt engineering optimization, both the scores and rationales more closely aligned with those of the human evaluator. The comments were precise and targeted, which can support student reflection and help them identify deficiencies for improvement, thereby promoting better learning outcomes and skills development.

Figure 7 Visual comparison of DeepSeek-V3's CPR assessment on the same student report, using different prompting strategies, alongside human teacher evaluation. (a) Human teacher evaluation; (b) LLM output with base prompt; (c) LLM output with PEG-Prompt; (d) LLM output with optimized PEG-Prompt (combined use of "Key content" and "Scoring examples"). Red words show segments with substantial deviations from human evaluations in scores or rationales. Green words highlights those with high consistency. CNN: convolutional neural network.

6 Conclusions

In this paper, we proposed use of a novel PEG-Prompt to enable objective and comprehensive assessment of CPR using LLMs. The approach works across six key dimensions, fully leveraging the Paul-Elder critical thinking framework alongside writing proficiency criteria and thereby capturing structural and linguistic features as well as reasoning, originality, and knowledge application. Additionally, we incorporated key content from student reports and representative examples of few-shot scoring into the PEG-Prompt to further boost CPR assessment performance. Extensive experiments on the PEG-CPR dataset demonstrated that PEG-Prompt enhances the alignment of LLM-generated scores with human evaluations, validating its effectiveness as a structured prompting strategy. Visualization analysis and qualitative case studies further validated the effectiveness of the proposed method in improving both the reliability and interpretability of LLM-generated assessments. Beyond methodological contributions, this work underscores the potential of PEG-Prompt to provide both valuable feedback for students and summative assessment in formal evaluation settings once appropriately calibrated with human markers.

Acknowledgments This work was supported in part by the Guangdong Provincial Teaching Quality and Teaching Reform Project (Grant No. SJZLGC202511), in part by the Southern University of Science and Technology Teaching Reform Project (Grant No. XJZLGC202414), and in part by the Medical and Health Science Program of Zhejiang Province (Grant No. WKJ-ZJ-26085).

Author Contributions Qingyang Sun contributed to original draft, review and editing, methodology, visualization, validation, software, project administration, and conceptualization; Jialu Zhang contributed to review and editing, investigation, and formal analysis; Peng Sheng contributed to software; Qianyi Wang contributed to visualization; Tianrui Wang contributed to validation; Heng Li contributed to supervision; Hanshu Zhan contributed to resources; Xiaoqing Zhang contributed to writing, review and editing, supervision, and funding acquisition; Jiang Liu contributed to supervision and funding acquisition. All authors whose names appear on the submission made substantial contributions to the conception or design of the work; approved the version to be published; and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflict of Interest The authors declare that they have no conflicts of interest.

Data Availability Statements The authors confirm that all data generated or analyzed during this study are included in this published article.

References

Amatriain, X. (2024). Prompt design and engineering: Introduction and advanced methods. *arXiv Preprint*, arXiv:2401.14423.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Cader, A. (2020). The potential for the use of deep neural networks in e-learning student evaluation with new data augmentation method. In: *Proceedings of the 21st International Conference on Artificial Intelligence in Education (AIED 2020)*. Ifrane: Springer, 37–42.

Cao, J. C., Li, M. Z. N., Wen, M., & Cheung, S. C. (2025). A study on prompt design, advantages and limitations of ChatGPT for deep learning program repair. *Automated Software Engineering*, *32*(1), 30.

Chen, J. Y., Wu, T., Ji, W., & Wu, F. (2024). WisdomBot: Tuning large language models with artificial intelligence knowledge. *Frontiers of Digital Education*, *1*(2), 159–170.

Dong, Z. A., Chen, J. Y., & Wu, F. (2025). LLM-driven cognitive diagnosis with SOLO taxonomy: A model-agnostic framework. *Frontiers of Digital Education*, *2*(2), 20.

Ennis, R. H. (1993). Critical thinking assessment. *Theory Into Practice*, *32*(3), 179–186.

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. Washington: American Philosophical Association.

Friesen, S., & Scott, D. (2013). *Inquiry-based learning: A review of the research literature*. Edmonton: Alberta Ministry of Education, 1–32.

Gemini Team, Google. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv Preprint*, arXiv:2403.05530.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., & et al. (2024). The Llama 3 herd of models. *arXiv Preprint*, arXiv:2407.21783.

Halpern, D. F. (2013). *Thought and knowledge: An introduction to critical thinking*. 5th ed. New York: Psychology Press.

Han, J., Yoo, H., Kim, Y., Myung, J., Kim, M., Lim, H., Kim, J., Lee, T. Y., Hong, H., Ahn, S. Y., & Oh, A. (2023). RECIPE: How to integrate ChatGPT into EFL writing education. In: *Proceedings of the Tenth ACM Conference on Learning @ Scale*. Copenhagen: ACM, 416–420.

Justice, C., Rice, J., Warry, W., Inglis, S., Miller, S., & Sammon, S. (2007). Inquiry in higher education: Reflections and directions on course design and teaching methods. *Innovative Higher Education*, *31*(4), 201–214.

Ke, Z. X., Inamdar, H., Lin, H., & Ng, V. (2019). Give me more feedback Ⅱ: Annotating thesis strength and related attributes in student essays. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence:

ACL, 3994–4004.

Lee, H. Y. (2014). Inquiry-based teaching in second and foreign language pedagogy. *Journal of Language Teaching and Research*, *5*(6), 1236–1244.

Li, S. J., & Ng, V. (2024). Automated essay scoring: A reflection on the state of the art. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami: ACL, 17876–17888.

Liu, A. X., Feng, B., Xue, B., Wang, B. X., Wu, B. C., Lu, C. D., Zhao, C. G., Deng, C. Q., Zhang, C. Y., Ruan, C., & et al. (2024a). DeepSeek-V3 technical report. *arXiv Preprint*, arXiv:2412.19437.

Liu, J., Nie, Q. S., Wang, X. Y., Xiao, Z. J., & Zhang, X. Q. (2025). Exploration of weakly supervised learning pedagogy under the background of generative artificial intelligence. *Computer Education*, (1), 198–203, 209. (in Chinese).

Liu, P. J., Zhang, W., Ding, Y. L., Zhang, X. F., & Yang, S. H. (2024b). SEMDR: A semantic-aware dual encoder model for legal judgment prediction with legal clue tracing. In: *Proceedings of the 2024 IEEE International Conference on Systems, Man, and Cybernetics*. Kuching: IEEE, 3447–3453.

Mansour, W. A., Albatarni, S., Eltanbouly, S., & Elsayed, T. (2024). Can large language models automatically score proficiency of written essays? In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. Torino: ACL, 2777–2786.

Mieg, H. A. (2019). *Inquiry-based learning–undergraduate research: The German multidisciplinary experience*. Cham: Springer.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050.

Nabhan, R. J. (2017). Integration of inquiry-based learning and ongoing assessment to develop English essay writing in upper intermediate level. *Open Journal of Modern Linguistics*, *7*(2), 90–107.

Page, E. B. (1967). Grading essays by computer: Progress report. In: *Proceedings of the Invitational Conference on Testing Problems*. Princeton: Educational Testing Service, 87–100.

Paul, R., & Elder, L. (2019). *The miniature guide to critical thinking concepts and tools*. Santa Rosa: The Foundation for Critical Thinking.

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., Manoli, C. C., Zacharia, Z. C., & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, *14*, 47–61.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527.

Stahl, M., Biermann, L., Nehring, A., & Wachsmuth, H. (2024). Exploring LLM prompting strategies for joint essay scoring and feedback generation. In: *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*. Mexico City: ACL, 283–298.

Sun, Y., Wang, S. H., Feng, S. K., Ding, S. Y., Pang, C., Shang, J. Y., Liu, J. X., Chen, X. Y., Zhao, Y. B., Lu, Y. X., & et al. (2021). ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv Preprint*, arXiv:2107.02137.

Sun, Q. Y., Wang, T. R, Sheng, P., Wang, Q. Y., Zhang, X. Q., & Liu, J. (2025). Exploration of intelligent evaluation methods for computer course project reports under the education intelligence framework. *Computer Education*, (1), 210–214. (in Chinese).

Waked, A., Pilotti, M., & Abdelsalam, H. M. (2024). Differences that matter: Inquiry-based learning approach to research writing instruction. *Scientific Reports*, *14*(1), 27941.

Wen, Q. S., Liang, J., Sierra, C., Luckin, R., Tong, R., Liu, Z. T., Cui, P., & Tang, J. L. (2024). AI for education (AI4EDU): Advancing personalized education with LLM and adaptive learning. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Barcelona: ACM, 6743–6744.

White, J., Fu, Q. C., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. In: *Proceedings of the 30th Conference on Pattern Languages of Programs*. Monticello: ACM, 5.

Xiao, C. R., Ma, W. X., Xu, S. X., Zhang, K. P., Wang, Y. F., & Fu, Q. (2024). From automation to augmentation: Large language models elevating essay scoring landscape. *arXiv Preprint*, arXiv:2401.06431.

Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. In: *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*. Toronto: ACL, 576–584.

Zhang, X. Q., Wei, J. Q., Ye, S. J., Xiao, Z. J., Hu, Y., Tang, J. G., & Liu, J. (2021). Multimedia meets archaeology: A novel interdisciplinary teaching approach. In: *Proceedings of the 2021 IEEE Frontiers in Education Conference*. Lincoln: IEEE, 1–8.