Computers & Education 243 (2026) 105511

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Leveraging prompt-based LLMs for automated scoring and feedback generation in higher education☆

Eman Mudhi AlGhamdi [a,c], Yuheng Li [b,c], Dragan Gašević [c], Guanliang Chen [c,*]

[a] College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia
[b] Department of Applied Social Sciences, The Hong Kong Polytechnic University, Hong Kong, China
[c] Faculty of Information Technology, Monash University, Melbourne, Australia

ARTICLE INFO

Keywords:
Automated essay scoring
Feedback generation
Prompt engineering
Learner-centered feedback
Higher education

ABSTRACT

As demand grows for personalized, scalable assessments in higher education (including both scoring and feedback provision), large language models (LLMs) have emerged as promising tools. While human educators typically perform scoring and feedback in a sequential and interrelated manner, existing research has largely addressed these tasks separately. This raises important questions about LLMs' ability to handle scoring and feedback within a single workflow and the extent to which task sequencing affects their performance. To address this gap, this study investigates how prompting LLMs to perform scoring and feedback either together in one single prompt (prompt composition) or separately in two consecutive prompts (prompt decomposition), and the order in which these tasks are prompted affect the performance of GPT-4o, a cutting-edge LLM, in postgraduate open-ended assessments. We analyzed the scoring performance across student groups of varying performance levels. To tailor GPT-4o-generated feedback to individual student learning needs, we embedded well-established learner-centered feedback principles into the prompt design and assessed the quality of the generated feedback based on these principles. The scoring results revealed that prompt effectiveness varied modestly across student groups, with higher scoring errors on lower quality submissions. In terms of generated feedback, GPT-4o demonstrated greater support for learner agency. Task order influenced how this agency was expressed: prompting feedback first fostered learner autonomy, while prompting it after scoring emphasized the student–teacher connection.

1. Introduction

As personalized learning becomes a central goal in education, there is a growing demand for educators to deliver timely, individualized assessment scores and feedback (Alsaiari et al., 2025). Meeting this demand has placed increasing pressure on instructors, particularly in writing-intensive courses (Xiao et al., 2025). In response, Automated Essay Scoring (AES) systems have emerged as scalable alternatives to manual grading, assigning rubric-aligned scores and generating formative feedback to support student improvement (Mizumoto & Eguchi, 2023; Xu et al., 2024). AES systems have shown strong assessment performance, achieving state-of-the-art results on widely used evaluation datasets such as ASAP (e.g., Cozma et al., 2018), TOEFL11 (e.g., Vajjala, 2018), and the AAE corpora (e.g., Ke et al., 2018), as discussed in the review by Ke and Ng (2019).

☆ This article is part of a Special issue entitled: 'GenAI enhanced learning' published in Computers & Education.

* Corresponding author.
E-mail addresses: emalghamdi@uj.edu.sa, eman.alghamdi91@gmail.com (E.M. AlGhamdi), yu-heng.li@polyu.edu.hk (Y. Li), dragan.gasevic@monash.edu (D. Gašević), guanliang.chen@monash.edu (G. Chen).

https://doi.org/10.1016/j.compedu.2025.105511
Received 12 July 2025; Received in revised form 19 November 2025; Accepted 19 November 2025
Available online 26 November 2025

Though generally effective, these AES systems often rely on machine learning, deep learning, and fine-tuned Large Language Models (LLMs) with inherent limitations. Traditional machine learning models depend on manually handcrafted features and struggle to capture semantic depth and contextual nuance (Xu et al., 2024). Deep learning reduces this reliance by automatically learning meaningful features from data, but it requires large annotated datasets and often struggles to maintain coherence in longer essays (Feng et al., 2024; Xu et al., 2024). While fine-tuning LLMs improves performance, it still demands significant computational resources and technical expertise, limiting broader educational adoption (Mayer et al., 2023).

Recent advances in generative LLMs, such as GPT-4, offer new opportunities for AES through prompt-based learning, which enables task execution via natural language instructions rather than model retraining or fine-tuning (Yu et al., 2025). This approach makes advanced AI tools more accessible for non-tech-savvy educators to apply in diverse assessment settings (An et al., 2025; Li et al., 2025; Liu et al., 2023; Mayer et al., 2023; Pack et al., 2024). Research has shown that prompt-based LLMs can match the scoring accuracy of state-of-the-art fine-tuned models (Chamieh et al., 2024) while also generating nuanced, human-like feedback (Masikisiki et al., 2023). However, most studies exploring prompt-based LLMs in higher education have tackled scoring and feedback generation as separate tasks, focusing on either scoring (e.g., Flodén, 2025) or feedback provision (e.g., Dai et al., 2023). Since human educators often perform scoring and feedback provision sequentially and interrelatedly (i.e., either by assigning a holistic score followed by feedback that explains the evaluation and offers suggestions for improvement, or by first drafting detailed feedback to identify strengths and areas for improvement, which then informs the assigned score), it is intuitive to investigate whether LLMs can effectively address both tasks and how the order of these tasks could influence their performance.

Driven by this, this study investigated how prompt-based LLMs handle scoring and feedback generation either jointly within a single prompt (prompt composition) or sequentially across two separate prompts (prompt decomposition), and how the order of these tasks (i.e., first scoring and then feedback generation, or the other way around) influences model performance. Prior research has shown that composing or decomposing these tasks, as well as their order, can affect LLM output (Jiang & Bosch, 2024; Stahl et al., 2024; Wu et al., 2024), though these effects remain inconclusive and may not generalize to open-ended writing tasks at the higher education level. Compared to primary and secondary education, such writing tasks pose greater grading challenges due to their increased conceptual complexity (Flodén, 2025). They also call for feedback that extends beyond score justification, fostering deeper reflection, active engagement, and personalization responsive to students' specific learning needs (Nicol & Macfarlane-Dick, 2006). This has led to interest in learner-centered feedback approach that guides future learning, highlights strengths and weaknesses, and fosters student ownership of the learning process (Aldino et al., 2025; Ryan et al., 2023). Grounded in these motivations, this study is guided by the following research questions:

- **RQ1:** How do (i) the number of tasks requested in a prompt (i.e., composition vs. decomposition); and (ii) the order of the tasks influence the performance of **LLM-based scoring** for open-ended essay tasks in higher education?
- **RQ2:** How do (i) the number of tasks requested in a prompt (i.e., composition vs. decomposition); and (ii) the order of the tasks influence the quality of **LLM-generated learner-centered feedback** for open-ended essay tasks in higher education?

We addressed the research questions using a dataset comprising 214 postgraduate project proposals from a data science course, each accompanied by instructor-assigned scores and commentary feedback grounded in standardized assessment rubrics. We designed four prompt configurations that varied in the number of tasks included (prompt composition vs. decomposition) and the order of tasks (scoring-first vs. feedback-first), using GPT-4o, a state-of-the-art generative AI model at the time of this research. To answer RQ1, we tackled automated scoring as a regression problem and evaluated GPT-4o's performance using both error-based metrics and rank-order consistency across high-performing, low-performing, and overall student groups. To address RQ2, we assessed the quality of GPT-4o-generated feedback based on its alignment with the learner-centered feedback framework. Specifically, we used a pre-trained classifier developed by Aldino et al. (2024) to automatically detect sentences reflecting the framework's dimensions: learner agency, sensemaking, and future impact. The findings of this study demonstrate how prompt design influences both scoring and the generation of personalized feedback, offering design implications for future adaptive, human-AI collaborative assessment systems.

2. Background

2.1. Automated Essay Scoring

Automated Essay Scoring (AES) has progressed from early statistical approaches (Page, 1966) to machine learning, deep learning techniques and, most recently, transformer-based large language models (Ke & Ng, 2019; Lagakis & Demetriadis, 2021; Xu et al., 2024). Generative Pre-trained Transformer (GPT) models introduced a new paradigm in AES by enabling scoring through prompt-based interaction, where natural language instructions guide model reasoning and evaluation (Pack et al., 2024). Among these models, GPT-4 has been identified as one of the most reliable models for automated grading, outperforming other prompt-based LLMs in reasoning and alignment with human educators (Chamieh et al., 2024; Lee et al., 2024; Pack et al., 2024). Its performance is further enhanced when augmented with scoring rubrics (Tang et al., 2024) and scoring examples (Chang & Ginter, 2024), even with minimal prompt engineering (Henkel et al., 2024). These findings underscore GPT-4's advanced reasoning capabilities and its contribution to improving automated assessment systems.

Despite advancements in scoring accuracy, current AES research remains heavily reliant on corpora such as ASAP[1] (Cozma et al., 2018), ICLE[2] (Persing & Ng, 2014), CLC-FCE[3] (Yannakoudakis & Briscoe, 2012), TOEFL11[4] (Vajjala, 2018), and AAE[5] (Ke et al., 2018), as consistently noted in reviews of the field (Ke & Ng, 2019; Lagakis & Demetriadis, 2021; Xu et al., 2024). These datasets primarily assess English proficiency, general writing skills, and argument structure, often within secondary or language-learning contexts, rather than focusing on the substantive quality of the written content that is essential in higher education. While some prompt-based LLM studies have begun to evaluate the quality of the written content in short-answer questions using public benchmark datasets (Jiang & Bosch, 2024) or undergraduate-level tasks (Morjaria et al., 2024), relatively little attention has been paid to open-ended essays in higher education. Such essays are typically longer, more complex, and discipline-specific, making them significantly more challenging for automated scoring systems (Flodén, 2025). Notably, focusing AES functionality solely on scoring accuracy is insufficient for educational contexts (Xu et al., 2024), as effective learning also depends on the provision of high-quality, personalized feedback that supports student understanding and improvement.

2.2. Automatic feedback generation

Feedback guides students in understanding their performance and progressing toward learning goals (Jensen et al., 2021; Nicol & Macfarlane-Dick, 2006). In educational systems, providing quality feedback is more effective in enhancing student learning outcomes than merely assigning scores (Deeva et al., 2021; Ke & Ng, 2019; Van der Kleij et al., 2015; Xu et al., 2024). Recent research highlights the growing potential of LLMs in feedback generation. For example, ChatGPT produces more detailed and readable feedback than teacher comments (Dai et al., 2023), and GPT-4 effectively supports student learning outcomes at a level comparable to human instructors (Escalante et al., 2023).

In the AES context, most LLM-generated feedback focuses on justifying predicted scores based on grading rubrics. These include studies by Jiang and Bosch (2024), Ormerod and Kwako (2024), Pack et al. (2024), Song et al. (2024), Tang et al. (2024), Xiao et al. (2025), and Feng et al. (2024). While Masikisiki et al. (2023) and Stahl et al. (2024) attempted to enhance feedback by prompting LLMs to suggest improvements for future work, their approaches remained narrowly focused on performance gaps and lacked the depth required to foster active learner engagement and ownership of the learning process. This reveals a gap in current research, in which feedback is often conceptualized as a one-way transmission of information focused primarily on identifying discrepancies between students' current performance and expected goals. Such conventional approaches marginalize learner agency by framing feedback as evaluative commentary rather than equipping students with self-regulation skills, which are essential for effective learning in higher education (Narciss et al., 2014; Nicol & Macfarlane-Dick, 2006; Schiller et al., 2025). In response, our study adopted a learner-centered feedback approach that emphasizes learner autonomy, active engagement, and actionable support for ongoing development.

The provision of quality feedback is typically framed as a learner-centered process that enables students to interpret feedback meaningfully and apply it in their subsequent tasks (Lin et al., 2023), promoting learner ownership and participation in the learning experience (Aldino et al., 2024). Ryan et al. (2023) proposed and validated the Learner-Centered Feedback framework grounded in established educational theory. Defined as guidance that helps students interpret, apply, and benefit from feedback, the framework is organized around three core dimensions, each encompassing a set of distinct components:

- **Future Impact:** Emphasizes how feedback supports students' future learning and development.
  - *Future Improvement:* Offers guidance for improving future work.
  - *Learning Outcomes and Skill Development:* Aligns feedback with broader academic and professional goals.
- **Sensemaking:** Helps students understand their current performance.
  - *Strengths and Weaknesses:* Identifies task-specific strengths and areas for improvement.
  - *Performance Summary:* Provides an overall evaluation relative to assessment criteria.
- **Agency:** Supports students in recognizing and taking charge of their own learning growth.
  - *Active Role:* Promotes active engagement with feedback and self-directed learning.
  - *Affirm and Encourage:* Recognizes and reinforces students' accomplishments.
  - *Strengthen Relationship:* Fosters a supportive connection between student and instructor.

[1] ASAP: Automated Student Assessment Prize.
[2] ICLE: International Corpus of Learner English.
[3] CLC-FCE: Cambridge Learner Corpus – First Certificate in English.
[4] TOEFL11: ETS Corpus of Non-Native Written English.
[5] AAE: Annotated Argumentative Essays.

Recent work by Aldino et al. (2024) demonstrated that the key dimensions of learner-centered feedback can be automatically identified in textual feedback using trained classifiers based on Natural Language Processing (NLP) techniques, highlighting the potential of such methods to assess alignment with the Learner-centered Feedback framework at scale. Building on this, our study shifts the focus from post-hoc evaluation to the generative process itself, investigating how prompt-based LLMs can be guided to produce personalized, pedagogically meaningful feedback for open-ended assessment tasks in higher education.

2.3. Prompt-based LLM approaches

Prompt-based learning leverages natural language instructions, known as prompts, to guide LLMs' task execution without the need for retraining (Masikisiki et al., 2023). For example, Xiao et al. (2025) demonstrated that GPT-4, when guided by well-crafted prompts, outperformed fine-tuned GPT-3.5 in scoring accuracy, underscoring the effectiveness of prompt-based learning.

Prompts can be deployed under various in-context learning settings, each differing in the amount of guidance provided to the model. Zero-shot learning supplies only task instructions, one-shot includes a single task example, and few-shot learning provide multiple examples (Mayer et al., 2023). Including examples generally improves scoring accuracy over zero-shot approaches (Feng et al., 2024; Henkel et al., 2024; Lee et al., 2024; Song et al., 2024), particularly with GPT-4 (Jiang & Bosch, 2024; Xiao et al., 2025). To ensure representativeness, shot selection typically includes samples from different score levels. However, adding more in-context examples does not always improve performance and may even hinder it due to the model's limited capacity to process longer input sequences effectively (Song et al., 2024).

Prompt design has emerged as a key factor in optimizing LLM-based AES systems (Feng et al., 2024; Liu et al., 2023; Masikisiki et al., 2023; Tang et al., 2024). Among the most effective prompt designs is the Chain-of-Thought (CoT) prompting, which encourages models to articulate reasoning steps before producing a final output (Kojima et al., 2022). AES studies have demonstrated that Zero-Shot CoT prompting enhances reasoning by allowing models to generalize across tasks without human-constructed step-by-step reasoning examples (Masikisiki et al., 2023; Stahl et al., 2024). Similarly, role-based prompting, which instructs the model to adopt a predefined role such as *"Teacher assistant"*, has been shown to enhance alignment with task expectations and improve the quality of both scoring and feedback (Chen et al., 2025; Song et al., 2024; Stahl et al., 2024).

Recent studies have shown that both the structure (i.e., prompt composition or decomposition) and the sequence of prompting scoring and feedback tasks affect LLM performance (e.g., Jiang & Bosch, 2024). In prompt composition, scoring and feedback are requested together in a single prompt (e.g., Stahl et al., 2024), whereas in prompt decomposition, the two tasks are separated across consecutive prompts within the same session (e.g., Ormerod & Kwako, 2024). Despite growing interest, prior findings offer no clear consensus on the optimal use of these prompt strategies or the order of the scoring and feedback provision tasks when working with LLMs. For example, Wu et al. (2024) argued that explanations are generally more accurate when generated after output prediction, as they directly reflect the model's reasoning. In contrast, Stahl et al. (2024) reported that generating explanations before scoring improved performance and prevented the model from justifying an inaccurate score. Adding further nuance, Jiang and Bosch (2024) found that the optimal sequence depends on prompt structure: scoring-first was more effective in composed prompts, whereas feedback-first yielded better results in decomposed formats. However, these findings were not validated within higher education context, making their generalizability to open-ended assessments in this context largely unknown. Driven by this gap, our study investigated how the number of tasks included in a single prompt and the order of these tasks affected LLM performance in scoring and feedback generation within higher education.

3. Methodology

3.1. Dataset

The dataset was sourced from a postgraduate-level course on introductory data science at an Australian University. Ethics approval was obtained from the university's Human Research Ethics Committee for using students' assessment submissions and their received feedback (Project Id: 29874).[6] As part of the course's assessment, students submitted a proposal for a data science project of their choice, comprising two sections: the Project Description, which outlined project goals and relevant data professionals' roles, and the Business Model underpinning the proposed data science project, which described the target beneficiaries, value proposition, and anticipated challenges.

In total, 401 student proposals were submitted in a single semester. Proposals were evaluated by experienced course instructors, who assigned a holistic numerical score ranging from 0 to 15 and provided textual feedback based on five rubric dimensions: (1) clarity of goals, (2) relevance to data science, (3) articulation of business value, (4) creativity, and (5) overall clarity. These instructor-assigned scores and rubric-based textual feedback served as the reference standard for evaluating and comparing the LLM-generated scores and feedback in order to address RQ1 and RQ2. Per the course assessment policy, these holistic scores were first scaled to the range [0, 100] and then mapped to five performance bands: 0–49 (N — Fail), 50–59 (P — Pass), 60–69 (C — Credit), 70–79 (D — Distinction), and 80–100 (HD — High Distinction). We removed all personally identifiable information during pre-processing of the collected project proposals to ensure student privacy. We also excluded non-content elements such as cover

[6] Data from this research will be made available by the corresponding author upon reasonable request.

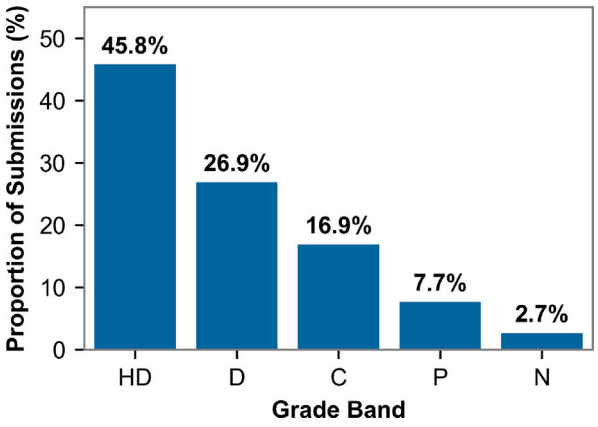

Fig. 1. Percentages of student proposals by grade band.

pages, tables of contents, headers, and footers, retaining only the proposal title, main body of text, and references (where available) as input to the LLM.

Due to the dataset's limited size and imbalanced distribution across grade bands (Fig. 1), we grouped student proposals into two performance categories based on human-assigned scores: High Performers (HD, D) and Low Performers (C, P, N), following established practices in educational research aimed at facilitating targeted analyzes and interventions tailored to group-specific learning needs (Farooq & Regnier, 2011; Jibeen & Khan, 2016; Kam & Umar, 2023). After randomly selecting four-shot exemplars as described in Section 3.3, we retained the remaining 107 Low Performer submissions and randomly sampled 107 proposals from the High Performer group, resulting in a balanced dataset of 214 proposals with an average word count of approximately 715 words. This sampling strategy was intended to support fair comparisons across proficiency levels while avoiding inflated performance metrics caused by class imbalance (Lee et al., 2024).

3.2. LLM selection

GPT-4 demonstrates superior scoring capabilities, surpassing earlier versions such as GPT-3.5, outperforming traditional models like SVM, and approaching the upper-bound performance of fine-tuned BERT (Chamieh et al., 2024). It also produces feedback closely aligned with human quality, outperforming open-source alternatives (Ziems et al., 2024). Its capacity to handle extended inputs (Masikisiki et al., 2023) makes it well-suited for evaluating long-form student submissions, such as those used in our study. For these reasons, we used GPT-4o, a more recent variant of GPT-4 that offers comparable performance, along with faster responses and lower API costs (OpenAI, 2024).

3.3. Prompt engineering

All prompts in this study follow the best practices established in prior research, such as those outlined by Chen et al. (2025) and Ziems et al. (2024). Chen et al. (2025) emphasizes several key elements that enable LLMs to generate high-quality responses:

- **Role Prompting:** The role of *"an experienced teacher in a master's-level data science course"* was assigned to the model to align its responses with the academic standards expected in higher education.
- **Chain-of-Thought Prompting:** Since our dataset lacks explicit reasoning steps for scoring and feedback generation, this study employs Zero-Shot CoT prompting. The phrase *"Let's think step by step"* was used as it has been shown to enhance LLM logic in solving various tasks (Chen et al., 2025; Kojima et al., 2022).
- **Few-Shot Prompting:** Two exemplar submissions from each performance category (high-performing and low-performing) were randomly selected from outside the 214-sample dataset used in this study, to mitigate the risk of data leakage (Chang & Ginter, 2024; Lee et al., 2024). Including more than two exemplars per category yielded minimal gains and introduced a risk of performance degradation due to excessive prompt length. To minimize ordering bias, the examples were presented in a random sequence (Song et al., 2024; Zhao et al., 2021).
- **Structured Formatting:** To ensure long text sections are processed as coherent units, triple quotes (''''''') were used to enclose the task description, scoring rubric, shots, and proposal being evaluated.

In addition, Ziems et al. (2024) further recommends LLM prompting guidelines that encourage adherence to structured output formats, addressing LLM-response issues noted by Pack et al. (2024):

- **Context Before Instruction:** Establishing contextual grounding before delivering instructions enhances the model's adherence to task-specific expectations.

- **Explicit Constraints:** Clear instructions, such as *"Even if uncertain, you must still select a score within this range"*, enforce response completeness.
- **Structured Output Specification:** We applied OpenAI's Structured Output enforcement rather than relying solely on natural language prompt instructions for output format specification.

To align the model's feedback with individual learner needs, the prompt included explicit instructions guiding the model to follow the learner-centered feedback framework, specifying its corresponding feedback dimensions (nam-ely learner agency, sensemaking, and future impact) and their associated components. The wording of the prompts used in this study was held constant to eliminate content variation as a confounding factor when comparing prompt composition and decomposition as well as task order (Masikisiki et al., 2023).

3.4. Experimental design

Given evidence that both the number of tasks in a prompt and the order in which they are presented can influence scoring accuracy and feedback quality (Jiang & Bosch, 2024; Stahl et al., 2024; Wu et al., 2024), we tested four prompting configurations that varied these two factors. The four experimental conditions were:

- **Composition-Scoring-First:** A single prompt first requests a holistic score (0–15), followed by feedback.
- **Composition-Feedback-First:** Feedback is generated before scoring within the same prompt.
- **Decomposition-Scoring-First:** Scoring and feedback are elicited in two separate consecutive prompts within the same session, with scoring first.
- **Decomposition-Feedback-First:** Feedback is generated first, followed by a separate prompt within the same session requesting a score based on that feedback.

To reduce variability in GPT-4o outputs, we ran each prompt five times per proposal and averaged the results (mean ± SD) (Flodén, 2025; Xiao et al., 2025). We fixed GPT-4o's parameters (temperature = 0, top-p = 0.01) to minimize randomness (Henkel et al., 2024; Jiang & Bosch, 2024) and processed each proposal in a separate API call to prevent context leakage (Flodén, 2025; Yancey et al., 2023). The Composition-Scoring-First prompt is presented in Fig. 2, annotated to reflect the prompt engineering practices detailed in Section 3.3, with other prompts provided in Appendix A.

3.5. Evaluation

3.5.1. Scoring evaluation

We addressed RQ1 by evaluating GPT-4o's scoring performance in terms of both the magnitude of scoring errors and the consistency of rank ordering relative to human-assigned scores. Specifically, Relative Merit Consensus (RMC) (Chang & Ginter, 2024) was used to assess whether the model preserved the correct rank order of student submissions based on their quality. RMC values range from 0 to 1, where higher values indicate stronger agreement between the model-generated and human-assigned rankings. This metric reflects perceived fairness in scoring, as students are more likely to view scores as unfair when lower-quality submissions receive equal or higher scores than their own. Scoring error was quantified using Mean Absolute Error (MAE), which captures the average deviation from human scores, and Root Mean Squared Error (RMSE), which places a greater penalty on larger discrepancies (Xu et al., 2024).

To statistically assess whether scoring errors measured by MAE and RMSE differed significantly between prompt pairs, we conducted two-sided paired permutation tests on absolute and squared errors using 10,000 resampled values. The null hypothesis assumed that there was no difference in scoring error between different prompt conditions. Permutation testing was chosen for its suitability with paired observations drawn from the same dataset (Riezler & Hagmann, 2024). It also provides greater statistical power than conventional parametric and nonparametric tests when applied to small samples with non-normal distributions (Dror et al., 2018), a distributional property confirmed in our data by Shapiro–Wilk tests and Q-Q plot results (Appendix B). To quantify the effect sizes of the observed scoring differences, we computed bias-corrected and accelerated (BCa) bootstrap confidence intervals around mean differences, following (Bestgen, 2022). Scoring performance was analyzed across the entire student population and separately for high- and low-performing groups, offering deeper insights into how the effectiveness of different prompt designs varies with the quality of student work and informing the development of adaptive, personalized, prompt-based LLM-powered assessment systems.

3.5.2. Feedback evaluation

To address RQ2, feedback quality was evaluated in terms of the presence of learner-centered components as defined in Ryan et al. (2023), enabling us to examine how different prompts influenced the pedagogical value of the generated feedback. Given the need for scalable evaluation over hundreds of feedback instances, we employed an automated classification approach using a BERT-based multi-label classifier developed by Aldino et al. (2024). This classifier was trained on 13,025 annotated feedback instances from 95 university courses and demonstrated strong alignment with human annotations, with accuracy and F1 scores above 0.90 across all learner-centered components, and Cohen's kappa values exceeding 0.80 for six out of eight components.

We applied the classifier at the sentence level to the LLM-generated feedback under each prompt, extracting binary predictions (present/absent) for each learner-centered component. For each prompt, we computed the proportion of feedback instances that included at least one sentence containing each component. To benchmark the LLM-generated feedback against human-written feedback, we also applied the classifier to the instructor comments included in our dataset.

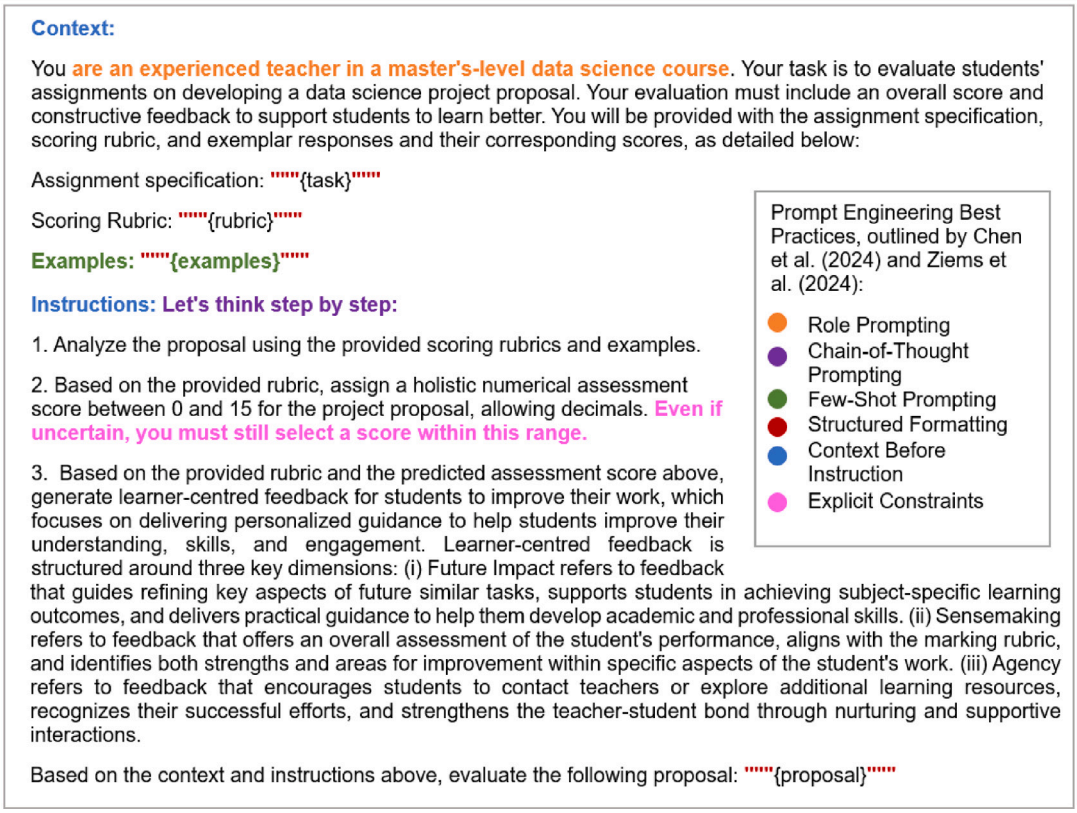

Fig. 2. Sample prompt from this study (Composition-Scoring-First), annotated to illustrate six prompt engineering strategies informed by Chen et al. (2025) and Ziems et al. (2024). Color-coded labels correspond to each prompting strategy.

Table 1
Scoring performance (mean ± standard deviation) across prompt configurations varying by number of tasks (composition vs. decomposition) and task order. Metrics are reported both across the entire student population (All) and separately for high-performing (High) and low-performing (Low) student groups. The best results in each column are in bold.

Tasks in a Prompt	Task Order	MAE (± SD)			RMSE (± SD)			RMC (± SD)		
		High	Low	All	High	Low	All	High	Low	All
Composition	Scoring-First	1.67 ± 0.03	**2.06** ± 0.02	1.86 ± 0.02	2.08 ± 0.03	**2.43** ± 0.03	2.26 ± 0.03	0.31 ± 0.01	0.42 ± 0.01	0.39 ± 0.01
	Feedback-First	1.59 ± 0.02	2.12 ± 0.06	**1.85** ± 0.02	1.98 ± 0.03	2.47 ± 0.05	**2.24** ± 0.02	**0.34** ± 0.01	**0.44** ± 0.01	**0.42** ± 0.01
Decomposition	Scoring-First	**1.49** ± 0.06	2.23 ± 0.09	1.86 ± 0.02	**1.85** ± 0.07	2.56 ± 0.08	**2.24** ± 0.02	**0.34** ± 0.01	0.43 ± 0.01	0.41 ± 0.01
	Feedback-First	1.66 ± 0.01	2.15 ± 0.01	1.91 ± 0.01	2.04 ± 0.01	2.52 ± 0.01	2.29 ± 0.01	0.31 ± 0.01	0.41 ± 0.00	0.39 ± 0.00

4. Results

4.1. Results on RQ1

4.1.1. Impact of prompt composition, decomposition, and task order on overall regression and ranking performance

Scoring results on all student submissions (Table 1) show that prompt composition and decomposition had a more pronounced impact under *Feedback-First* conditions. *Composition-Feedback-First* achieved the best scoring accuracy, demonstrating a 3.1% lower MAE, 2.2% lower RMSE, and 7.7% higher RMC relative to *Decomposition-Feedback-First*, which produced the weakest results. Statistical testing (Table 2) confirms that the RMSE difference between these two prompts reached statistical significance ($p = .046$), although the effect size was modest, with a 95% confidence interval for the error difference of [–0.14, –0.01]. These findings suggest that placing feedback before scoring in the prompt sequence may reduce larger scoring errors by priming the model with a more informed interpretation of the response.

4.1.2. Impact of prompt composition, decomposition, and task order across student performance levels

Across ranking and error results, GPT-4o's scoring performance varied substantially by student proficiency level (Table 1). High-performing submissions consistently received more accurate scores, with 19%–34% lower MAE, and 14.4–27.7% lower RMSE than

Table 2
Permutation test results of RMSE differences between prompt pairs across student performance groups. Each cell shows the RMSE difference calculated as (row prompt–column prompt), along with 95% BCa bootstrap confidence intervals. Positive values indicate higher RMSE for the prompt in the row. Prompt abbreviations: C-SF = Composition-Scoring-First, C-FF = Composition-Feedback-First, D-SF = Decomposition-Scoring-First, D-FF = Decomposition-Feedback-First. Significance: $p < .05$ (*), $p < .01$ (**), $p < .001$ (***).

Group	Prompt	C-FF	D-SF	D-FF
High Performers	C-SF	*** 0.11 [0.06, 0.17]	*** 0.23 [0.14, 0.33]	0.04 [−0.07, 0.16]
	C-FF	–	** 0.12 [0.05, 0.20]	−0.07 [−0.19, 0.04]
	D-SF	–	–	*** −0.19 [−0.31, −0.10]
Low Performers	C-SF	−0.04 [−0.11, 0.04]	*** −0.13 [−0.23, −0.07]	* −0.10 [−0.21, −0.03]
	C-FF	–	* −0.10 [−0.19, −0.02]	−0.06 [−0.15, 0.00]
	D-SF	–	–	0.03 [−0.03, 0.12]
Overall	C-SF	0.03 [−0.02, 0.08]	0.03 [−0.04, 0.09]	−0.04 [−0.11, 0.03]
	C-FF	–	−0.002 [−0.07, 0.06]	* −0.07 [−0.14, −0.01]
	D-SF	–	–	−0.07 [−0.13, −0.00]

Table 3
Permutation test results of MAE differences between prompt pairs across student performance groups. Each cell shows the MAE difference calculated as (row prompt–column prompt), along with 95% BCa bootstrap confidence intervals. Positive values indicate higher MAE for the prompt in the row. Prompt abbreviations: C-SF = Composition-Scoring-First, C-FF = Composition-Feedback-First, D-SF = Decomposition-Scoring-First, D-FF = Decomposition-Feedback-First. Significance: $p < .05$ (*), $p < .01$ (**), $p < .001$ (***).

Group	Prompt	C-FF	D-SF	D-FF
High Performers	C-SF	*** 0.09 [0.05, 0.15]	*** 0.18 [0.11, 0.28]	0.01 [−0.08, 0.10]
	C-FF	–	* 0.09 [0.03, 0.17]	−0.08 [−0.18, 0.01]
	D-SF	–	–	*** −0.17 [−0.27, −0.09]
Low Performers	C-SF	−0.07 [−0.14, 0.02]	*** −0.17 [−0.26, −0.08]	* −0.10 [−0.20, −0.03]
	C-FF	–	* −0.10 [−0.19, −0.01]	−0.03 [−0.13, 0.05]
	D-SF	–	–	0.07 [−0.01, 0.15]
Overall	C-SF	0.01 [−0.04, 0.06]	0.01 [−0.06, 0.08]	−0.05 [−0.11, 0.02]
	C-FF	–	−0.002 [−0.06, 0.06]	−0.06 [−0.12, 0.01]
	D-SF	–	–	−0.05 [−0.12, 0.00]

those of low-performing students. This reflects stronger alignment with human-assigned scores in terms of absolute accuracy for high performers. However, despite lower absolute accuracy, higher RMC values for low-performing submissions suggest that the model more reliably preserved relative rank order among weaker submissions than among stronger ones. This reveals a performance asymmetry: GPT-4o assigned more precise scores to stronger submissions but exhibited greater consistency in ranking weaker ones.

In addition, scoring error patterns across prompts revealed a trade-off in evaluation metrics between student groups: gains for one group were often accompanied by declines in the other. For example, *Decomposition-Scoring-First* achieved the strongest results for high performers (MAE = 1.49, RMSE = 1.85) but underperformed for low performers (MAE = 2.23, RMSE = 2.56). In contrast, *Composition-Feedback-First* yielded more balanced outcomes across groups (MAE: 1.59 vs. 2.12; RMSE: 1.98 vs. 2.47). This suggests that certain prompts may affect scoring accuracy differently for high- and low-performing students.

To further explore these group-level differences, we examined MAE and RMSE across all prompt configurations within high- and low-performing student groups. Tables 1–3 suggest that prompt configuration may influence scoring accuracy differently across student performance groups. Among high performers, *Decomposition-Scoring-First* yielded the strongest results, achieving the lowest MAE (1.49), and the lowest RMSE (1.85). Statistical comparisons (Tables 2 and 3) further support this trend, showing that *Decomposition-Scoring-First* significantly outperformed all other prompts. Specifically, it reduced MAE by 0.18 ($p < .001$) and RMSE by 0.23 ($p < .001$) compared to *Composition-Scoring-First*, which exhibited the lowest scoring performance among all prompts. In contrast, for low-performing submissions, the *Composition-Scoring-First* emerged as the most effective, achieving the lowest MAE (2.06) and RMSE (2.43) among all prompts (Table 1). It significantly outperformed *Decomposition-Scoring-First* ($p < .001$) and *Decomposition-Feedback-First* ($p < .05$) in reducing prediction errors (Tables 2 and 3). These consistent findings across regression and statistical analyses provide strong evidence of prompt–proficiency interactions, whereby the most effective prompting configuration varies by student performance group. However, the relatively small effect sizes suggest that while these interactions are statistically robust, their practical impact may be limited.

4.2. Results on RQ2

To address RQ2, we examined how the number of tasks in one prompt and task order shaped the quality of GPT-4o-generated feedback, as measured by the presence of feedback components defined in the Learner-centered Feedback framework (Ryan et al., 2023). As shown in Fig. 3, GPT-4o included agency-supportive elements — such as *Active Role* and *Strengthen Relationship* — at higher rates than the human-crafted feedback in our dataset. As shown in Fig. 3, *Composition-Scoring-First* and *Decomposition-Scoring-First* prompts resulted in the highest inclusion of the *Strengthen Relationship* component (66.4% and 66.8%, respectively), suggesting that

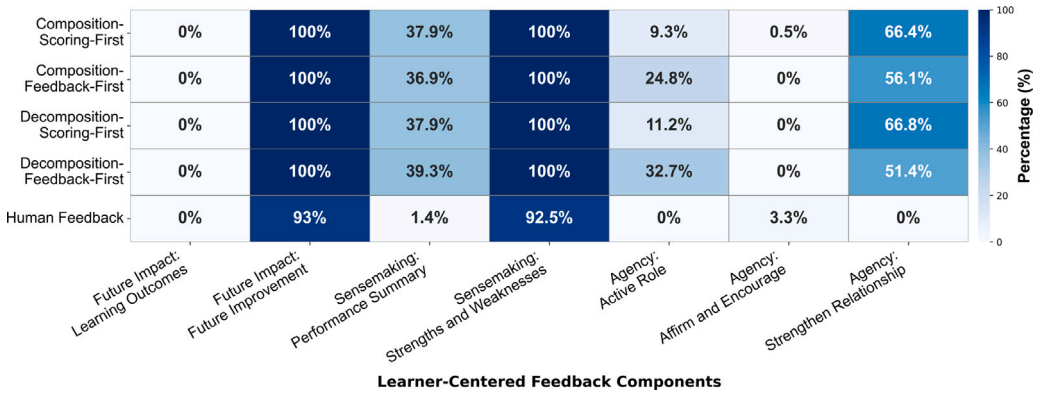

Fig. 3. Percentage of feedback instances (n = 214) that included different learner-centered feedback component across the prompt settings and the human-crafted feedback.

Scoring-First prompts encourage the model to adopt a more relational and supportive tone. In contrast, *Feedback-First* prompts were more effective at fostering learner autonomy. Specifically, the *Active Role* component appeared most frequently in *Decomposition-Feedback-First* (32.7%), followed by *Composition-Feedback-First* (24.8%), indicating that generating feedback before scoring primes the model to emphasize student autonomy and engagement. These patterns underscore the influence of task order on the model's ability to express learner agency, particularly through the *Strengthen Relationship* and *Active Role* components.

Beyond task order effects on agency, other learner-centered components in LLM-generated feedback showed minimal variation across prompts. All prompts led the model to consistently include *Future Improvement* suggestions and identification of *Strengths and Weaknesses* (100%). However, the *Performance Summary* component was present in just 37%–39% of feedback instances across prompts. Despite this, GPT-4o demonstrated this component more frequently than human educators in our dataset, which almost omitted it (1.4%). However, this strength was not mirrored in other dimensions, as GPT-4o and human feedback consistently lacked affirmation and learning outcome alignment. Specifically, the *Affirm and Encourage* component was almost entirely missing from model-generated responses (< 1%) and was present in only 3.3% of human-crafted feedback, highlighting a shared limitation in delivering emotionally supportive feedback. Similarly, the *Learning Outcomes* component was absent across all GPT-generated and human-crafted feedback. This absence indicates a shared gap in how both AI and human graders deliver emotionally supportive and learning-aligned feedback, pointing to areas for future improvement in addressing these underrepresented components.

5. Discussion

5.1. Prompt-based LLM approaches for scoring

GPT-4o consistently scored high-performing submissions with greater accuracy, though this may stem from a tendency to assign higher scores broadly, rather than an improvement in evaluative precision. This pattern aligns with concerns in the literature that LLM-based scoring systems tend to assign overly generous scores to open-ended essays, particularly in higher education (Chang & Ginter, 2024; Flodén, 2025). One likely explanation is that GPT models struggle to differentiate quality in longer, content-based student responses, as noted by Chang and Ginter (2024), who found that such responses often received uniformly high scores regardless of quality. To further explore this possibility, we conducted an additional correlation analysis between essay length and assigned scores. The results revealed a consistent positive relationship across various conditions: for human scores, the Spearman correlation with essay length was approximately 0.25 ($p < .001$), while it ranged from 0.20 to 0.28 ($p < .01$) for GPT-assigned scores. These findings reinforce prior work showing that text length influences the scoring of both human educators and automated systems, with longer essays often receiving higher scores (Fleckenstein et al., 2020). Such similarity is to be expected, given that LLM-based scoring models are trained on human data and may therefore inherit human judgment biases. As Fleckenstein et al. (2020) noted, overscoring tendencies linked to text length may sometimes function as a heuristic but also risk becoming a source of bias when length is rewarded at the expense of quality. In addition, we performed a sanity check by pairing student submissions that received the same human-assigned score (thus controlling for quality) but differed in length. We found that GPT scores for longer submissions had a higher mean absolute error (MAE = 1.92) than shorter submissions (MAE = 1.71), although a paired t-test confirmed that this difference was not statistically significant. These results underscore the importance of examining how text length shapes GPT-4o's scoring patterns and of exploring mitigation strategies — such as explicitly instructing LLMs to prioritize quality over length — to reduce potential bias in higher education contexts.

Considering all student submissions regardless of their quality, the *Composition-Feedback-First* prompt yielded slightly better overall regression performance than all other prompts, including a statistically significant improvement in RMSE over *Decomposition-Feedback-First*. Interestingly, this outcome diverges from prior findings by Jiang and Bosch (2024), who reported that feedback-first prompts were more effective under decomposition-based prompting in the context of short-answer questions from secondary

education, using the ASAP dataset. The discrepancy between our findings and prior work suggests that optimal prompt design may depend not only on task composition and order but also on the complexity of student responses being evaluated. Further research is warranted to examine this interaction and determine how prompt design can be adapted to diverse response types.

A more nuanced insight emerges when scoring performance is analyzed by student performance, our results revealed that not all prompts were equally effective with their success varying slightly by the performance level of the student work being assessed. High-performing submissions were most accurately scored and ranked using *Decomposition-Scoring-First* prompt, outperforming all other prompt settings across multiple evaluation metrics. This suggests that initiating the prompt with a scoring task is sufficient for the model to make accurate judgments when evaluating well-structured responses typically produced by high-performing students. However, the same prompt yielded the lowest-scoring performance for low-performing submissions. This contrast indicates that *Decomposition-Scoring-First* apparent effectiveness may stem from the model's difficulty in detecting lower-quality responses when scoring is prompted in isolation, leading the model to over-predict. For low-performing students, the *Composition-Scoring-First* prompt was the most effective, indicating that integrating feedback and scoring instructions within a unified prompt may offer essential contextual cues that enable the model to better interpret and assess underdeveloped responses.

Taken together, these findings support the growing body of evidence that the number of tasks included in a prompt and their order affect the reliability of LLM-based scoring (Jiang & Bosch, 2024; Stahl et al., 2024; Wu et al., 2024). However, our results show that the effect is not uniform and varies depending on the quality of the student's response, supporting the presence of prompt–proficiency interactions. While several of these effects reached statistical significance, the corresponding effect sizes and overall magnitude of improvement remained relatively small. Our findings suggest that although prompt composition, decomposition, and task order influence model behavior, they are insufficient on their own to ensure reliable scoring across diverse student performance levels.

Moreover, although the analysis by performance level uncovered several statistically significant differences between prompt configurations, the aggregated results across all submissions identified only one significant improvement, with *Composition-Feedback-First* outperforming *Decomposition-Feedback-First* in RMSE. This contrast underscores a key methodological implication: aggregate evaluations may mask important differences between student performance groups, reinforcing the value of disaggregated analysis for ensuring fair and accurate assessment across diverse groups of students. Recognizing this, our findings suggest that performance-aware prompting could serve as a guiding principle for future AI-based assessment systems, in which prompts are not only optimized for task characteristics but also dynamically adapted based on real-time features of student submissions. Such systems could leverage surface-level and semantic cues — such as linguistic complexity, coherence, or rhetorical structure — to infer response characteristics and tailor prompt configurations accordingly. For example, submissions with fragmented or underdeveloped linguistic features may benefit from composed prompts that elicit scoring and feedback jointly, offering richer contextual grounding, whereas well-structured responses may be more accurately evaluated using decomposed prompts that separate these tasks. This direction aligns with recent advances such as AdaptPrompt (Wang et al., 2025), which dynamically adjusts prompts based on task characteristics—for instance, by extracting and integrating implicit textual relationships between sentence pairs to guide adaptive prompt construction, enhance the model's interpretive focus, and improve prediction accuracy.

Furthermore, given the varying predictive performance of different prompting strategies across student performance levels, future AI-based assessment systems may benefit from more advanced prompting techniques. One promising direction is ensemble-style prompting (Zhang et al., 2024), where multiple prompts are used to assess the same response and their outputs are aggregated (e.g., via majority voting or weighted averaging) to improve robustness and mitigate prompt-specific biases. In addition, systems could prompt LLMs to indicate their confidence levels in generated scores, enabling more cautious and informed use (Detommaso et al., 2024). For example, low-confidence predictions might be flagged for instructor review or presented with cautionary indicators to students. Complementarily, prompting LLMs to provide scoring rationales — that is, justifications for their assignment scores — could enhance model transparency and allow students to critically evaluate the trustworthiness and instructional value of AI-generated outputs.

5.2. Prompt-based LLM approaches for feedback generation

The influence of prompt design was evident in feedback components related to learner agency. Specifically, the *Active Role* and *Strengthen Relationship* dimensions were observed exclusively in LLM-generated feedback, in contrast to human educators, with their presence shaped by the sequencing of tasks within the prompt. The *Feedback-First* configuration, which prompts the model to reflect on the student's work before assigning a score, prioritizes attention to the learner's needs and areas for growth before judgment. This task order more frequently elicited feedback that encourages learner initiative, such as: *"Feel free to discuss any questions or seek guidance on areas you find challenging"*, aligning with the *Active Role* component of learner agency. Conversely, in the *Scoring-First* configuration, the model begins by evaluating the student's performance, often priming the output toward justification of the assigned score or recognition of effort. Feedback in this setup tended to include statements such as: *"Remember, your efforts are recognized, and I encourage you to continue building on this strong foundation"*, reflecting the *Strengthen Relationship* component. These differences indicate that task order shapes the model's communicative orientation and can be strategically leveraged to adapt generated feedback to specific pedagogical goals or learner needs.

Across all prompts, LLM-generated feedback consistently reflected key learner-centered components, particularly *Strengths and Weaknesses* and *Future Impact*, aligning with the pedagogical view that effective feedback should highlight both achievements and areas for improvement (Nicol & Macfarlane-Dick, 2006), while also guiding postgraduate students in applying complex concepts to practice (Lin et al., 2023). Also, GPT-4o produced more comprehensive summaries of student work within the *Performance*

Summary component than human educators, aligning with Dai et al. (2023), who noted that ChatGPT tends to produce more coherent summaries of student performance than human instructors. The LLM's tendency to restate student ideas in a structured and detailed manner may help clarify key points and support deeper understanding, particularly in complex postgraduate tasks.

The *Affirmation and Encouragement* component was limited in both LLM-generated and human-crafted feedback. Among all GPT-4o outputs, only one brief sentence (*"your work is appreciated"*) was explicitly identified as affirmational. Human feedback, though also infrequent, included similar concise and direct affirmations such as *"Good work"* or *"Well done otherwise!"*. Upon closer examination, affirmational language in GPT-4o's feedback was occasionally present but was predominantly embedded within performance-oriented evaluations rather than expressed as standalone affective statements. For example, the sentence *"You've done a commendable job in detailing the roles and responsibilities of data scientists, which shows a clear understanding of the project's scope"* conveys a positive tone; however, its primary communicative function is evaluative, focused on task performance. As a result, such feedback is more likely to be classified under components like *Performance Summary* or *Strengths and Weaknesses* rather than *Affirm and Encourage*. This pattern reflects GPT-4o's general tendency to prioritize task-relevant commentary over explicit emotional support. Nevertheless, prior research suggests that when positive feedback is framed in terms of students' task performance, it can still enhance motivation (Ryan et al., 2023), indicating that GPT-4o's task-focused praise may offer motivational benefits even when affective intent is implicit.

While our findings indicate that GenAI-generated feedback — when appropriately prompted — can display several components of learner-centered feedback, it is important to clarify that this should not be interpreted as evidence that LLMs are inherently better at providing such feedback than human educators. Well-prepared and conscientious teachers are highly capable of delivering nuanced, personalized, and context-sensitive feedback that supports learner autonomy and growth. In fact, many of the learner-centered features observed in GenAI feedback stem from prompt instructions designed by humans and from training data that encode human values and pedagogical strategies. Therefore, GenAI's ability to produce such feedback is ultimately contingent on human guidance, both during pretraining and deployment. Any perceived alignment with learner-centered principles in GenAI-generated feedback depends heavily on how well such principles are explicitly embedded into the prompts and task design. Moreover, these findings are based on a relatively small dataset collected from a single postgraduate-level course. As such, they should not be overgeneralized to suggest that GenAI outperforms well-trained educators. Still, our findings highlight the potential of GPT-4o to serve as a valuable assistant in supporting learner-centered feedback practices, particularly when integrated into a human–AI collaborative workflow. For example, GenAI could generate initial feedback that is then refined or augmented by teachers, offering a scalable and practical solution for delivering personalized feedback, enhancing learner engagement, and reducing feedback-related workload for educators (Morjaria et al., 2024). This, in turn, enables teachers to focus more on instructional activities that require their professional creativity and complex pedagogical reasoning (Flodén, 2025).

6. Limitations and future work

This study acknowledges several limitations that also point to opportunities for future research. First, the dataset used in this study is relatively small, having been collected from a single postgraduate-level course. While it yields valuable insights into the effects of different prompt configurations on automated scoring and feedback provision, the findings may be context-specific — shaped by the subject matter and academic level of the course — and thus may not generalize to other educational settings without further validation. Moreover, although the observed effects were statistically significant, the limited sample size necessitated a binary classification of students into high and low performers. This constraint hindered the ability to detect more fine-grained variations in prompt effectiveness across narrower grade bands, as the reduced number of samples within each band diminished the statistical power of the analysis. Second, although feedback was assessed from the learner-centered perspective, the evaluation focused solely on component presence, overlooking broader contextual and ethical factors such as clarity, relevance, truthfulness (i.e., the accuracy of model-generated content, avoiding hallucinated or fabricated statements) and the potential impact on student trust (Wu et al., 2024). To overcome these limitations, future research should leverage larger and more diverse datasets to enhance the robustness and generalizability of the findings. Richer datasets would also support more nuanced analyzes of prompt effectiveness on both scoring and feedback generation, and facilitate investigations into model behaviors such as overprediction tendencies and sensitivity to variations in writing quality in complex open-ended assessments. Prompts should also be refined to enhance the pedagogical quality of LLM-generated feedback by incorporating illustrative examples aligned with each learner-centered component. Finally, broader feedback evaluation frameworks should be adopted to move beyond the presence of learner-centered components and assess additional factors of feedback quality.

7. Conclusion

This study investigated how the number of tasks requested in a prompt (i.e., composition vs. decomposition) and task order influence scoring reliability and the quality of AI-generated feedback using GPT-4o. Scoring results indicated modest variation in prompt effectiveness across performance levels: while *Decomposition-Scoring-First* prompts performed slightly better for high-performing submissions, *Composition-Scoring-First* prompts were comparatively more effective for lower-performing ones. Regarding feedback provision, LLM-generated feedback analyzed in this study showed higher alignment with most of the learner-centered principles than feedback provided by human educators, with task order influencing the model's communicative stance: *Feedback-First* prompts encouraged active learner engagement, whereas *Scoring-First* prompts emphasized strengthening teacher-student relationship. These findings highlight GPT-4o's potential to assist instructors by supporting learner-centered education through personalized feedback, while offering design implications for adaptive, prompt-based assessment systems that adapt scoring prompts to students' proficiency levels and tailor feedback prompts to their agency needs.

CRediT authorship contribution statement

Eman Mudhi AlGhamdi: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yuheng Li:** Writing – review & editing, Supervision, Project administration. **Dragan Gašević:** Writing – review & editing. **Guanliang Chen:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author used ChatGPT (OpenAI, GPT-4o) to improve readability and flow of the manuscript. After using this tool, the author reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was conducted as part of Eman Mudhi AlGhamdi's Master's thesis at Monash University. The degree was sponsored by a scholarship from the University of Jeddah, supported by the Kingdom of Saudi Arabia's Ministry of Education. The sponsor had no involvement in the study design, data collection, analysis or interpretation, writing of the manuscript, or the decision to submit the article for publication. This research was also in part funded by the Australian Research Council (DP220101209; DP240100069) and Jacobs Foundation (CELLA 2 CERES).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compedu.2025.105511.

Data availability

Data will be made available on request.

References

Aldino, A. A., Tsai, Y.-S., Gupte, S., Henderson, M., Nath, D., Gašević, D. ., & Chen, G. (2025). Analytics of learner-centered feedback: A large-scale case study in higher education. *Computers & Education*, *237*, Article 105360.

Aldino, A. A., Tsai, Y. S., Mello, R. F., Gašević, D., & Chen, G. (2024). Enhancing feedback quality at scale: Leveraging machine learning for learner-centered feedback. *Computers and Education: Artificial Intelligence*, *7*, Article 100332. http://dx.doi.org/10.1016/j.caeai.2024.100332.

Alsaiari, O., Baghaei, N., Lahza, H., Lodge, J. M., Boden, M., & Khosravi, H. (2025). Emotionally enriched AI-generated feedback: Supporting student well-being without compromising learning. *Computers & Education*, Article 105363.

An, S., Zhang, S., Guo, T., Lu, S., Zhang, W., & Cai, Z. (2025). Impacts of generative AI on student teachers' task performance and collaborative knowledge construction process in mind mapping-based collaborative environment. *Computers & Education*, *227*, Article 105227.

Bestgen, Y. (2022). Please, don't forget the difference and the confidence interval when seeking for the state-of-the-art status. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 5956–5962). Marseille, France: European Language Resources Association.

Chamieh, I., Zesch, T., & Giebermann, K. (2024). Llms in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In *Proceedings of the 19th workshop on innovative use of nlp for building educational applications (bea 2024)* (pp. 309–315). https://aclanthology.org/2024.bea-1.25/.

Chang, L. H., & Ginter, F. (2024). Automatic short answer grading for finnish with chatgpt. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*, 23173–23181. http://dx.doi.org/10.1609/aaai.v38i21.30363.

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patterns*, http://dx.doi.org/10.1016/j.patter.2025.101260.

Cozma, M., Butnaru, A., & Ionescu, R. T. (2018). Automated essay scoring with string kernels and word embeddings. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: short papers)* (pp. 503–509). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P18-2080.

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y. S., & Gašević, G. (2023). Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE international conference on advanced learning technologies* (pp. 323–325). IEEE, http://dx.doi.org/10.1109/ICALT58122.2023.00100.

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, *162*, Article 104094.

Detommaso, G., Bertran, M. A., Fogliato, R., & Roth, A. (2024). Multicalibration for confidence scoring in LLMs. *International Conference on Machine Learning*, *1062*, 4–10641.

Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1383–1392). http://dx.doi.org/10.18653/v1/P18-1128.

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*, 57. http://dx.doi.org/10.1186/s41239-023-00425-2.

Farooq, M. S., & Regnier, J. C. (2011). Role of learning styles in the quality of learning at different levels. *Informatica Economica*, *15*.

Feng, H., Du, S., Zhu, G., Zou, Y., Phua, P. B., Feng, Y., Zhong, H., Shen, Z., & Liu, S. (2024). Leveraging large language models for automated chinese essay scoring. In *International conference on artificial intelligence in education* (pp. 454–467). Springer, http://dx.doi.org/10.1007/978-3-031-64302-6_32.

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S., & Köller, O. (2020). Is a long essay always a good essay? the effect of text length on writing assessment. *Frontiers in Psychology*, http://dx.doi.org/10.3389/fpsyg.2020.562462.

Flodén, J. (2025). Grading exams using large language models: A comparison between human and ai grading of exams in higher education using chatgpt. *British Educational Research Journal*, *51*, 201–224. http://dx.doi.org/10.1002/berj.4069.

Henkel, O., Hills, L., Boxer, A., Roberts, B., & Levonian, Z. (2024). Can large language models make the grade? an empirical study evaluating llms ability to mark short answer questions in k-12 education. In *Proceedings of the eleventh ACM conference on learning@ scale* (pp. 300–304). http://dx.doi.org/10.1145/3657604.3664693.

Jensen, L. X., Bearman, M., & Boud, D. (2021). Understanding feedback in online learning–A critical review and metaphor analysis. *Computers & Education*, *173*, Article 104271.

Jiang, L., & Bosch, N. (2024). Short answer scoring with gpt-4. In *Proceedings of the eleventh ACM conference on learning@ scale* (pp. 438–442). http://dx.doi.org/10.1145/3657604.3664685.

Jibeen, T., & Khan, M. A. (2016). Development of an academic achievement risk assessment scale for undergraduates: Low, medium and high achievers. *Multidisciplinary Journal of Educational Research*, *6*, 23–50. http://dx.doi.org/10.17583/remie.2016.1697.

Kam, A. H., & Umar, I. N. (2023). Would gamification affect high and low achievers differently? a study on the moderating effects of academic achievement level. *Education and Information Technologies*, *28*, 8075–8095. http://dx.doi.org/10.1007/s10639-022-11519-1.

Ke, Z., Carlile, W., Gurrapadi, N., & Ng, V. (2018). Learning to give feedback: Modeling attributes affecting argument persuasiveness in student essays. In *IJCAI* (pp. 4130–4136). https://abs/10.5555/3304222.3304344.

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI-19* (pp. 6300–6308). http://dx.doi.org/10.24963/ijcai.2019/879.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In *Advances in neural information processing systems* (pp. 22199–22213). http://dx.doi.org/10.5555/3600270.3601883.

Lagakis, P., & Demetriadis, S. (2021). Automated essay scoring: A review of the field. In *2021 international conference on computer, information and telecommunication systems* (pp. 1–6). IEEE, http://dx.doi.org/10.1109/CITS52676.2021.9618476.

Lee, G. G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, Article 100213. http://dx.doi.org/10.1016/j.caeai.2024.100213.

Li, Y., Raković, M., Srivastava, N., Li, X., Guan, Q., Gašević, D., & Chen, G. (2025). Can AI support human grading? Examining machine attention and confidence in short answer scoring. *Computers & Education*, *228*, Article 105244.

Lin, J., Dai, W., Lim, L. A., Tsai, Y. S., Mello, R. F., Khosravi, H., Gasevic, D., & Chen, G. (2023). Learner-centred analytics of feedback content in higher education. In *LAK23: 13th international learning analytics and knowledge conference* (pp. 100–110). http://dx.doi.org/10.1145/3576050.3576064.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*, 1–35. http://dx.doi.org/10.1145/3560815.

Masikisiki, B., Marivate, V., & Hlophe, Y. (2023). Investigating the efficacy of large language models in reflective assessment methods through chain of thought prompting. In *Proceedings of the 4th african human computer interaction conference* (pp. 44–49). http://dx.doi.org/10.1145/3628096.3628747.

Mayer, C. W., Ludwig, S., & Brandt, S. (2023). Prompt text classifications with transformer models! an exemplary introduction to prompt-based learning with large language models. *Journal of Research on Technology in Education*, *55*, 125–141. http://dx.doi.org/10.1080/15391523.2022.2142872.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*, Article 100050. http://dx.doi.org/10.1016/j.rmal.2023.100050.

Morjaria, L., Burns, L., Bracken, K., Levinson, A. J., Ngo, Q. N., Lee, M., & Sibbald, M. (2024). Examining the efficacy of chatgpt in marking short-answer assessments in an undergraduate medical program. *International Medical Education*, *3*, 32–43. http://dx.doi.org/10.3390/ime3010004.

Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Goguadze, G., & Melis, E. (2014). Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education*, *71*, 56–76.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*, 199–218. http://dx.doi.org/10.1080/03075070600572090.

OpenAI (2024). Introducing gpt-4o and more tools to chatgpt free users. Retrieved from https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/, (Accessed 10 2025).

Ormerod, C., & Kwako, A. (2024). Automated text scoring in the age of generative ai for the gpu-poor. *English Journal of Educational Measurement and Evaluation*, *5*, http://dx.doi.org/10.59863/OKUU1904.

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, *6*, Article 100234. http://dx.doi.org/10.1016/j.caeai.2024.100234.

Page, E. B. (1966). The imminence of. grading essays by computer. *The Phi Delta Kappan*, *47*, 238–243, http://www.jstor.org/stable/20371545.

Persing, I., & Ng, V. (2014). Modeling prompt adherence in student essays. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1534–1543). http://dx.doi.org/10.3115/v1/P14-1144.

Riezler, S., & Hagmann, M. (2024). *Validity, reliability, and significance: empirical methods for NLP and data science*. Springer Nature, http://dx.doi.org/10.1007/978-3-031-57065-0.

Ryan, T., Henderson, M., Ryan, K., & Kennedy, G. (2023). Identifying the components of effective learner-centred feedback information. *Teaching in Higher Education*, *28*, 1565–1582. http://dx.doi.org/10.1080/13562517.2021.1913723.

Schiller, R., Fleckenstein, J., Höft, L., Horbach, A., & Meyer, J. (2025). On the role of engagement in automated feedback effectiveness: Insights from keystroke logging. *Computers & Education*, Article 105386.

Song, Y., Zhu, Q., Wang, H., & Zheng, Q. (2024). Automated essay scoring and revising based on open-source large language models. *IEEE Transactions on Learning Technologies*, http://dx.doi.org/10.1109/TLT.2024.3396873.

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 (BEA 2024)* (pp. 283–298). Mexico City, Mexico: Association for Computational Linguistics, https://aclanthology.org/2024.bea-1.23/.

Tang, X., Chen, H., Lin, D., & Li, K. (2024). Harnessing llms for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon*, *10*, http://dx.doi.org/10.1016/j.heliyon.2024.e34262.

Vajjala, S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, *28*, 79–105. http://dx.doi.org/10.1007/s40593-017-0142-3.

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, *85*, 475–511. http://dx.doi.org/10.3102/0034654314564881.

Wang, B., Wang, Z., Xiang, W., & Mo, Y. (2025). Adaptive prompt learning with distilled connective knowledge for implicit discourse relation recognition. *IEEE Transactions on Audio, Speech and Language Processing*, http://dx.doi.org/10.1109/TASLPRO.2025.3527135.

Wu, X., Zhao, H., Zhu, Y., Shi, Y., Yang, F., Hu, L., Liu, T., Zhai, X., Yao, W., Li, J., et al. (2024). Usable xai: 10 strategies towards exploiting explainability in the llm era. http://dx.doi.org/10.48550/arXiv.2403.08946, arXiv preprint arXiv:2403.08946.

Xiao, C., Ma, W., Song, Q., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2025). Human-ai collaborative essay scoring: A dual-process framework with llms. In *Proceedings of the 15th international learning analytics and knowledge conference* (pp. 293–305). http://dx.doi.org/10.1145/3706468.3706507.

Xu, W., Mahmud, R., & Hoo, W. L. (2024). A systematic literature review: Are automated essay scoring systems competent in real-life education scenarios? *IEEE Access*, *12*, 77639–77657. http://dx.doi.org/10.1109/ACCESS.2024.3399163.

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 (BEA 2023)* (pp. 576–584). http://dx.doi.org/10.18653/v1/2023.bea-1.49.

Yannakoudakis, H., & Briscoe, T. (2012). Modeling coherence in esol learner texts. In *Proceedings of the seventh workshop on building educational applications using NLP* (pp. 33–43). https://aclanthology.org/W12-2004/.

Yu, S., Androsov, A., & Yan, H. (2025). Exploring the prospects of multimodal large language models for automated emotion recognition in education: Insights from Gemini. *Computers & Education*, *232*, Article 105307.

Zhang, C., Liu, L., Wang, C., Sun, X., Wang, H., Wang, J., & Cai, M. (2024). Prefer: Prompt ensemble learning via feedback-reflect-refine. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(17), 19525–19532.

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning* (pp. 12697–12706). PMLR, https://proceedings.mlr.press/v139/zhao21c.html.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, *50*, 237–291. http://dx.doi.org/10.1162/coli_a_00502.