# Open-ended Structured Question Assessment with Human-LLM Collaboration

ANONYMOUS AUTHOR(S)
SUBMISSION ID: 1380

Open-ended structured questions (OSQs) are valuable for assessing both factual knowledge and reasoning, but grading them is labor-intensive due to diverse and ambiguous responses. Large language models (LLMs) offer automation potential, yet existing LLM-driven grading methods often lack transparency, handle nuance poorly, and minimize instructor oversight, limiting their reliability in real educational contexts. We present *VeriGrader*, an instructor–LLM collaborative grading system designed to address these challenges. On the LLM side, *VeriGrader* segments responses by scoring points and incorporates instructor feedback into iterative grading through in-context learning of LLM. On the instructor side, *VeriGrader* allows instructors to visually and interactively review, refine, and annotate model outputs, thereby preserving transparency and agency. We validate *VeriGrader* through a usage scenario and a user study with instructors, showing it improves grading consistency, reduces workload, and fosters trust, highlighting the benefits of human–AI collaboration for rigorous and transparent assessment.

CCS Concepts: • **Applied computing** → **Education**; • **Human-centered computing** → **Graphical user interfaces**; • **Information systems** → *Information retrieval*.

Additional Key Words and Phrases: LLMs, Education

**ACM Reference Format:**
Anonymous Author(s). 2025. Open-ended Structured Question Assessment with Human-LLM Collaboration. In *Proceedings of The ACM CHI conference on Human Factors in Computing Systems (ACM CHI 26)*. ACM, New York, NY, USA, 27 pages. https://doi.org/10.1145/XXXXXXX.XXXXXXX

## 1 Introduction

**O**pen-ended **S**tructured **Q**uestion (OSQ) represents an important form of educational assessment. Unlike fill-in-the-blank or multiple-choice questions, they not only evaluate students' mastery of knowledge but also assess their ability to organize language and reason logically [5, 24]. As a simple example (Figure 1-A), consider the OSQ: "Why should we brush our teeth every day?" To answer this question, students are expected to understand the functions of tooth brushing, namely cavity prevention and fresh breath, and to articulate these functions in a well-organized manner (Figure 1-B). The score of each response depends on the number of scoring points it matches (Figure 1-C). A response such as "Brushing removes food bits so we don't get holes in our teeth" (Figure 1-D) emphasizes cavity prevention, while "It makes your mouth clean and your breath not smell bad" (Figure 1-E) emphasizes fresh breath. Each addresses only part of the expected content, whereas "Brushing keeps teeth healthy and makes your breath fresh" (Figure 1-F) represents a more complete response.

Grading OSQ requires extensive domain expertise and teaching experience, creating substantial workload challenges for instructors [32, 35]. The core difficulty in grading OSQ stems from their inherently multidimensional, semi-structured, and highly diverse nature. Instructors must holistically evaluate the correctness, logical coherence, argumentative depth,


Manuscript submitted to ACM

and language clarity of student responses, often customizing grading rubrics to specific questions. Taking Figure 1 as an example again, students may give other responses, such as "My mom asked me to" or "I'm not embarrassed to talk to people up close," are more ambiguous and difficult to evaluate. In such cases, it is hard to determine the correctness and how the response aligns with cavity prevention and fresh breath. If determined, instructors must reconcile diverse responses into a unified and fair assessment. In large-scale classes or under tight deadlines, instructors face significant pressure to balance consistency, efficiency, and fairness, often leading to repeated reviews and corrections of prior grading decisions.

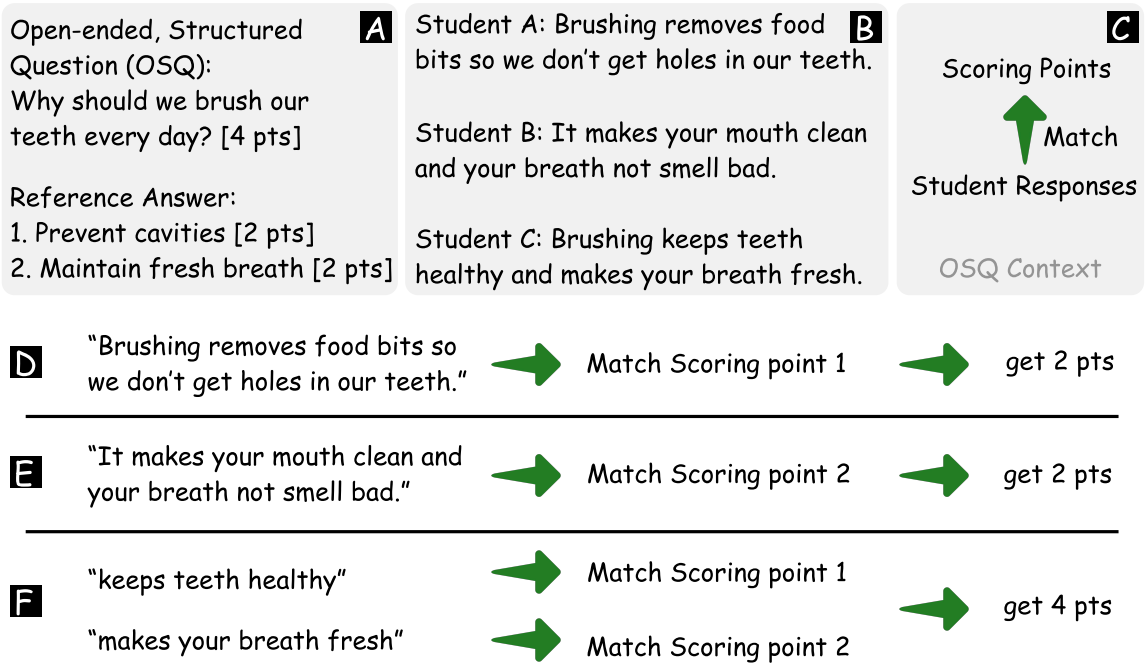


Fig. 1. (A) An OSQ and its reference answer with two scoring points. (B) Student responses. (C) Grading process. (D, E, F) Grading results, where the first two responses address only one scoring point each, whereas the third response covers two scoring points.

Recent advances in large language models (LLMs) have demonstrated impressive capabilities in natural language processing (NLP), motivating growing interest in their use for educational assessment [14, 29, 31, 50]. However, notable limitations remain when addressing subjective assessments, particularly in open-ended structured questions (OSQs), which often involve highly semantic, context-dependent, and nuanced reasoning. Most existing automated grading methods operate as "black-box" processes: they output only a single score while concealing why points were assigned and which parts of the reference answer supported them [6, 46, 49]. Such opaqueness falls short of the transparency and explainability required in grading contexts. Furthermore, LLMs continue to struggle with ambiguity [36] and context-dependent complexity in student responses [8], often resulting in inaccurate judgments. These challenges highlight the necessity of instructor involvement to guide and calibrate automated grading.

The interviews with stakeholders also argue that preserving instructor agency in grading is crucial for the widespread adoption and trust of AI-assisted grading systems in real educational environments. Without efficient and transparent

human-AI collaboration mechanisms, model errors may remain undetected or unresolved, ultimately undermining student trust in feedback and fostering educator resistance towards AI grading solutions. Relevant research also showed that when providing AI support, it is crucial to give users control and display instant previews of AI results [27].

In this study, we propose *VeriGrader*, an instructor-LLM collaborative grading system. *VeriGrader* is designed to address the aforementioned challenges: (1) ensuring grading consistency and reliability and (2) supporting transparency and controllability in human–AI collaboration. On the LLM side, we design a prompt-engineering framework that not only segments and scores student responses according to the scoring points but also incorporates instructors' feedback on LLM-based grading results into subsequent rounds of automated grading through in-context learning. On the instructor side, we develop an interactive, visual interface that allows instructors to intuitively validate grading results by visually examining how student responses align with the scoring points, and to annotate those they consider incorrect in order to provide feedback. We present a representative usage scenario to demonstrate *VeriGrader*'s utility. Besides, we conduct a user study to evaluate *VeriGrader*'s effectiveness. The results suggest that by fostering an instructor-LLM collaboration, *VeriGrader* largely reduces the grading workload and improves scoring accuracy. The collected user feedback offers valuable insights into AI-based educational applications, and we have shared them with peer researchers.

In sum, our contributions can be summarized as follows:

- **Problem Characterization:** We conduct stakeholder interviews and analyze graded exam papers to distill the workflow and requirements of human–AI collaboration for intelligent assessment of open-ended structured question (OSQ).
- **System Design and Implementation:** We develop a prototype system, *VeriGrader*, that supports instructor-LLM collaborative grading of OSQ. *VeriGrader* enables instructors to visually review the grading results produced by the LLM, interactively refine them, and provide feedback to the LLM to guide improved grading performance.
- **Evaluation:** Empirical validation of *VeriGrader* in authentic educational contexts, demonstrating its advantages in grading efficiency and instructor agency, alongside an exploration of future potential and directions for human-AI collaboration in education.

## 2 Related Work

### 2.1 Traditional Educational Assessment

Educational assessment spans a wide spectrum of practices, from objective assessment like multiple-choice and true/false questions to subjective assessment such as essays, open-ended problem solving, and creative assignments.

Objective assessments are prized for their efficiency, scalability, and reliable scoring, making them well-suited for testing factual knowledge and basic comprehension. However, their one-right-answer format inherently limits the ability to assess complex cognitive skills and higher-order understanding [7, 38]. In contrast, subjective assessments are crucial for gauging higher-order learning outcomes such as critical thinking, analytical reasoning, and creative expression [3, 38]. They require students to synthesize information and construct coherent arguments, and they also provide opportunities to demonstrate deep conceptual understanding, capabilities that are difficult to capture through purely objective formats.

Purely manual approaches to evaluating subjective assessments face substantial challenges. First, the workload and time demands are considerable. During high-stakes periods such as final examinations, instructors may be required to grade hundreds of submissions under tight deadlines, resulting in significant time pressure [11]. Second, scoring

consistency is difficult to maintain. Even experienced instructors may assign variable or even erroneous scores to identical responses due to factors such as fatigue, mood fluctuations, or differing interpretations of the scoring rubric [16, 23, 34]. The dual issues of high grading workload and variability in scoring accuracy have long constrained manual educational assessment. These inherent limitations have therefore motivated extensive research into AI-assisted approaches to automated assessment [2, 14, 22, 31, 45].

This study focuses on a common type of assessment item, which we term open-ended structured questions (OSQs). OSQs permit open responses but still require students to organize their answers with a certain degree of structure. They inherently combine objective criteria, such as factual correctness, with subjective judgments, such as clarity of reasoning and coherence of expression, thereby enabling a more comprehensive evaluation of both knowledge mastery and higher-order cognitive skills.

### 2.2 LLMs in Educational Assessment

LLM-as-Judge leverages large language models as automated evaluators to assess the quality, correctness, and preference of text, code, or multimodal content, thereby reducing human evaluation costs and enhancing efficiency in research, education, and system optimization [12]. This study particularly focuses on educational assessment.

In educational assessment, LLMs have been applied across diverse grading scenarios, including short-answer evaluation [14, 29], multiple-choice grading [31, 50], essay or report scoring [5, 22, 37, 45], and mathematical problem assessment [2, 44]. Unlike traditional rule-based systems that rely on keyword matching or shallow surface features, LLMs can accommodate varied expression styles and capture higher-order reasoning that was previously difficult to evaluate automatically. Grading typically presents a cold-start problem, as newly encountered questions typically lack sufficient labeled samples for reference. To mitigate this limitation, few-shot learning can leverage representative student responses [4], rubric injection can embed explicit scoring rubrics into prompts [21], and chain-of-thought reasoning can structure model outputs into step-by-step evaluations [43]. These approaches have all been effective to some extent in improving the grading accuracy of LLM. In certain settings, such as programming assignments, LLMs have achieved performance on par with human instructors [13, 20, 42].

Despite these advances, real-world adoption remains limited. Early systems [28, 46, 49] typically returned only a final numerical score without explanatory reasoning, thereby constraining instructor agency. Later efforts introduced reason generation and explanation interfaces [6, 46], but interaction largely remained one-directional: while instructors could review explanations, systems were unable to incorporate corrections or adapt grading rubrics. Furthermore, to the best of our knowledge, no prior work has systematically examined LLM-assisted grading for OSQs.

We propose a collaborative OSQ grading framework and develop a prototype system that goes beyond static scoring and post-hoc explanation. Tailored to the characteristics of OSQs, our approach interprets student responses in natural language and aligns them with the intended scoring points. It further integrates scoring with explicit explanation and incorporates instructor feedback into an iterative grading loop, thereby accelerating the grading process while embedding instructors' pedagogical intent and professional judgment.

### 2.3 Human-AI Collaboration

In parallel with rapid advances in AI capabilities, a substantial body of HCI research has examined how to integrate human expertise into AI-driven workflows to improve reliability, accountability, efficiency, and calibrated trust in real-world decision making [1, 9, 15, 19, 41].

Prior work commonly distinguishes two dominant interaction paradigms [33]: **AI-assisted decision making** and **human-supervised AI**. These paradigms are differentiated by who holds the default decision authority and when AI acts within the workflow. In AI-assisted decision making, the AI functions as an advisor, providing nonbinding recommendations, supporting rationales, or uncertainty estimates, while the human retains authority over the final decision and the AI does not act autonomously. In education, LLMs have been used to generate adaptive lesson content [10] and to suggest code explanations or modifications [18, 25], which instructors or students can selectively adopt. In contrast, human-supervised AI generates a provisional default decision (e.g., a score), which is then subject to human review, ratification, or override. This paradigm emphasizes efficiency while maintaining human oversight through post-hoc verification. In educational settings, most automated scoring systems adopt this paradigm, where the AI produces a provisional score that instructors may review or override. Subsequent systems have introduced enhanced supervisory features, such as rationales, uncertainty flags, or alternative scores [6, 21, 46], yet these remain constrained to a static, post-hoc review model in which the AI's decision logic does not adapt based on instructor feedback.

In complex tasks, humans and AI tend to assume more equal roles, forming a paradigm of mutual assistance and collaborative work [17, 26, 39, 48]. AI is not limited to providing recommendations but can also participate directly in decision-making. Human users are not merely reviewers; their feedback from evaluations can be incorporated into the system to further enhance the performance of AI. For example, LightVA [47], a lightweight visual analysis framework, employs LLM agents that remain accessible throughout the entire analysis process, enabling users to generate appropriate visualizations and obtain interpretations of findings through iterative conversational interaction. Smartboard [26] interprets multimodal status information from basketball games, incorporates the coach's tactical intentions, provides tactical recommendations, and ultimately supports collaborative tactic development with the coach.

Current educational assessment systems either emphasize efficiency through human-supervised AI or maintain instructional control via AI-assisted decision-making, but seldom integrate both. We propose a human-AI collaborative grading framework that combines transparent reasoning with bidirectional instructor–AI interaction and communication, allowing the system to iteratively align with instructors' professional judgment while preserving the efficiency of automation.

## 3 Background Knowledge

This section introduces the background knowledge of our work.

**Open-ended structured question (OSQ)** is a distinct assessment category positioned between fully open-ended questions (e.g., fully free-form essays) and fully structured questions (e.g., multiple-choice questions). Students compose their responses to OSQ freely based on their individual understanding of the questions and their linguistic habits. The instructor assesses OSQ responses based on a reference answer, represented as a list of scoring points with assigned score values.

**Scoring points** serve as the fundamental units for converting qualitative student responses into quantitative scores, ensuring consistent and fine-grained assessment. Each scoring point represents a logically independent and semantically complete knowledge unit that can be individually identified, assessed, and scored. The student response should cover as many predefined scoring points as possible. Due to the different linguistic habits and understanding of the questions, the same scoring point can be covered by various expressions, posing challenges for instructors to assess the student responses.

Figure 1 presents a concrete example using a simple OSQ "Why should we brush our teeth every day? [4']" Two scoring points may be: (1) preventing cavities [2'] and (2) keeping breath fresh [2']. One student may respond "Brushing

removes food bits so we don't get holes in our teeth," which addresses the first scoring point. Another student may write "It makes your mouth clean and your breath not smell bad," which corresponds to the second scoring point. Some responses may even cover both points, such as "Brushing keeps teeth healthy and makes your breath fresh."

## 4 Informing Interface Designs

We first conducted preliminary interviews to ensure that the system design aligns with the practical requirements of real-world teaching contexts. Based on the interviews, we compile design requirements that guide the visual and interaction design of the interface.

### 4.1 Preliminary interviews

**Interviewees.** Interviewees include (1) three high school instructors (T1-3), who, on average, have more than eight years of experience in grading and are potential users, (2) one teaching assistant (TA) who recently graded hundreds of exam papers, and (3) one academic affairs administrator (ADM), who frequently deals with students' requests for fairness and transparency in grades.

**Procedure.** The interview was conducted in a one-on-one manner, each lasting approximately 40 minutes. All participants provided informed consent, and their statements and opinions were used for academic research purposes. The interview was structured around four themes, addressed in sequence: the grading considerations, current grading workflow, challenges in grading, and potential improvement strategies.

**Results.** The interview results are summarized as follows:

- **Grading considerations.** ADM emphasized three essential factors in the grading process: fairness, accuracy, and interoperability. Fairness requires that all test papers be evaluated according to the same rubrics. Accuracy entails not overlooking correct responses, not misjudging incorrect responses as correct, and ensuring that partial scores are properly summed to yield the correct total. Interpretability means that both the total grade and its components, i.e., partial scores of scoring points, must be justified, with clear explanations of why the score was awarded or withheld. Importantly, the entire process must adhere to procedural justice. In extreme cases, when a student challenges or appeals the score, the instructor must be able to provide sufficient, well-reasoned explanations; otherwise, such situations could constitute a serious teaching incident.
- **Current grading workflow.** Interviews with T1-3 and TA revealed typical practices for grading OSQ. The grading process begins with evaluating a single response. First, the instructor identifies the relevant scoring points for the OSQ from the student response. Next, these scoring points are assessed individually, with partial scores assigned as appropriate. The assigned scores are then summed to produce the total grade. Once individual response has been graded, the process proceeds to grading a batch of responses, where the same rubrics are applied consistently across the set. After this, the scoring rubrics may be refined based on emerging patterns or ambiguities identified during grading. Finally, a double-check is performed across responses to ensure consistency, fairness, and accuracy in the overall grading process.
- **Challenges in grading.** After T1-3 and TA shared their grading workflow, we further inquired about the challenges they encounter during this process. A recurring issue is the presence of diverse expressions in student responses: although two responses may convey the same underlying idea, differences in wording often complicate consistent evaluation. Another challenge is ambiguous expressions, where student responses are phrased in ways that make it difficult to determine whether they should be considered correct or incorrect.

Finally, instructors noted the tediousness of double-checking, as they often need to review multiple responses to ensure that similar responses receive consistent treatment, which significantly increases the grading workload.

- **Potential improvement strategies.** Finally, we asked whether they perceived any areas for improvement, particularly from the perspectives of artificial intelligence and human-computer interaction. In response, T1, T3, and TA highlighted the following potential solutions. From the **artificial intelligence** perspective, T1, T3, and TA agree that, with the recent progress of large language models (LLMs), using automated grading is a good option, especially since LLMs have already achieved human-level performance in textual understanding. However, they are also concerned about ensuring accuracy and fairness. From the **human-computer interaction** perspective, the current grading process is indeed inconvenient and places a cognitive burden on instructors, such as memorizing scoring points for grading and frequently switching between multiple papers. Furthermore, given the uncertainty inherent in LLMs, the use of user interfaces to facilitate human intervention is even more crucial. For instance, T3 imaged an instructor being able to toggle an annotation label with a single click or automatically updating the rationale when a label is modified.

### 4.2 Empirical Analysis of Graded OSQs

We hope to derive insights from past graded exam papers, particularly regarding how instructors interpret student responses and determine scores. Thus, we analyze 141 graded exam papers containing open-ended, structured questions (OSQs). The grading annotations observed on these OSQs were primarily centered around the scoring points. Through manual examination, the meaning of these annotations can be categorized into the following categories.

**Correct:** Responses that align with the scoring point, either verbatim or semantically equivalent. Factually accurate and well-reasoned alternatives not in the reference answer are also considered correct. Such responses demonstrate accurate conceptual understanding. In the graded papers, correct responses are usually indicated with a check mark placed near the corresponding text.

**Wrong:** Responses that contradict the reference content or contain factual errors. Such responses reveal misunderstanding or misinterpretation. These are typically marked with a cross mark next to the relevant text.

**Unclear:** Responses whose meaning cannot be precisely judged against the scoring points in the reference answer due to ambiguity. Common cases include vague wording, incomplete or confused concepts, missing reasoning steps, or partial overlap with multiple scoring points. These are often underlined or indicated by question marks.

These annotations indicate that, after identifying potential scoring points within the responses, the instructors were further required to classify them into three categories and assign the corresponding scores.

## 5 *VeriGrader* System Overview

This section presents the detailed design of *VeriGrader*.

### 5.1 Design Rationales

Manual grading of OSQ is inherently time-consuming and repetitive. LLMs offer the potential to improve efficiency by understanding the responses and thereby automating the grading. However, fully automated, black-box grading is unsuitable in this context; instead, a hybrid workflow is required in which automated outputs are reviewed and confirmed by instructors, and the feedback of instructors, in turn, steers the model towards better grading. Such a design retains human authority while enabling efficiency gains, ensuring that ambiguous or borderline cases receive careful attention. Therefore, we further derive the following design rationales.

**R1: Receiving fine-grained and interpretable feedback.** Effective grading systems must provide more than a single aggregate score. Fine-grained and interpretable feedback is essential for clarifying where points were awarded or deducted and for explaining how specific response components align with scoring points. Such structured feedback ensures accuracy, fairness, and interpretability of the grading.

**R2: Supporting learnable, adaptable scoring behavior.** The scoring outputs generated by LLMs are not always accurate or fully aligned with the instructor's grading intent. To address this limitation, the system must provide mechanisms that allow instructors' feedback on scores and explanations to be incorporated. Establishing such a feedback loop ensures that refinements made by instructors can iteratively inform subsequent model behavior, thereby enabling the system to progressively adapt to and reflect instructors' grading philosophies.

**R3: Providing intuitive visual hints of grading results.** Effective grading requires that annotation and scoring results be conveyed in a clear and readily interpretable form. Beyond numerical scores, the system should visually encode the judgments from LLM or instructors directly on the response text, for example, through question marks for unclear text. Such visual hints make the rationale behind scores transparent, help instructors and students quickly locate supporting evidence, and foster a shared understanding of grading outcomes.

**R4: Integrating flexible user interactions to steer the grading.** To be effective, the user interface must support intuitive and flexible interactions with LLMs. Rather than requiring instructors to explicitly craft prompts, the system should enable implicit prompt construction through direct and visually guided interactions. In addition, instructors need the ability to review and compare alternative responses, and to clearly align each scoring point with the corresponding reference rubrics. Such interaction capabilities not only reduce cognitive burden but also ensure that instructors can efficiently validate model outputs and maintain consistency in grading.

### 5.2 Workflow

*VeriGrader* follows a multistage architecture and supports an iterative workflow (Figure 2). Given a set of student responses and a reference answer, the system first tries to segment the student response into pieces, maps these pieces to the scoring points, and categorizes them into three predefined categories (Figure 2-A). The segmentation, categorization, and grading reasons are mainly visualized in three panels, allowing the instructors to review and explain whether the student responses are correctly graded (Figure 2-B). During the review process, the instructor can refine the segmentation, categorization, and grading reasons as needed and finally confirm them (Figure 2-C). The confirmed grading results can be used as few-shot exemplars for subsequent rounds.

In particular, after the LLM completes automated grading, the instructor can review a small set of representative student responses to verify that response pieces are correctly categorized, aligned with reference scoring points, and supported by accurate reasons. Verified cases can be marked as high-quality exemplars, guiding subsequent iterations of grading. By iteratively repeating these steps, the system progressively refines the grading quality, after which instructors perform final grading on the student responses and export the corrected outputs.

### 5.3 Grading with LLM

Following the actual workflow of instructors, we design the grading process to ensure that each decision is both interpretable and verifiable. To support this, we incorporate a chain-of-thought strategy that encourages the model to "think step by step." Specifically, the LLM aims to segment the student response into non-overlapping response pieces based on the scoring points and classify each piece into **Correct**, **Wrong**, or **Unclear** based on the scoring points while generating reasons. From the output of grading each response, we can obtain multiple quadruples <**response piece**,

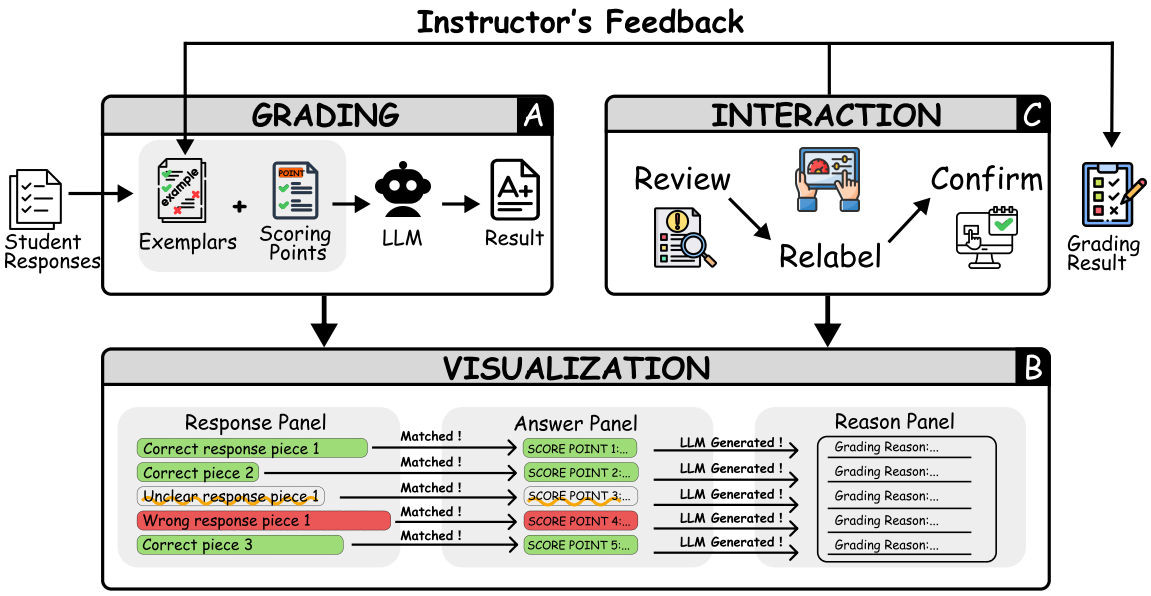


Fig. 2. *VeriGrader* workflow. (A) After upload, the system assembles a grading prompt from the question, reference answer, and exemplars, and calls the LLM to grade student responses. (B) The grading results of the LLM are visually exposed to the instructor in three panels. (C) Instructors review the grading results, relabel them as needed, and finally confirm them. The confirmed grading results can be used as few-shot exemplars for subsequent rounds.

**scoring point**, **category**, **reason**>, and further credit the total score for the response. This design makes reasoning transparent at the fine-grained level while ensuring that the final output remains standardized and consistent with the required schema. To operationalize this workflow, we design a structured grading prompt that clearly specifies the task and expected output format (Figure 3).

To further guide the LLM's reasoning and ensure standardized output, the system also supports the inclusion of a small number of exemplars provided by instructors during the collaborative grading process for few-shot learning. They are organized into two complementary categories, involving fine-grained guidance to macro-level demonstration:

- **point-level exemplars.** Each point-level exemplar is essentially a quadruple in the format <response piece, scoring point, category, reason>. They regulate the LLM's reasoning at a fine-grained level by illustrating how a specific response piece should be classified with respect to predefined scoring points. Unlike the task description in the prompt, which remains abstract, these exemplars explicitly address how the concrete response piece is matched with the matched scoring point, and explain whether it deserves a score or not.
  **response-level exemplars.** Each response-level exemplar consists of multiple quadruples with the aforementioned structure, all extracted from the same student response. Each exemplar is presented as a complete structured output in the same JSON format required from the model, facilitating the LLM's understanding of the distinctions among scoring points and the mapping between response pieces and scoring points. They operate at a macro level, providing complete grading cases that have been validated by instructors.

The two types of exemplars provide complementary guidance. Point-level exemplars shape local reasoning and argumentation at the response piece level, whereas response-level exemplars reinforce structural consistency and global

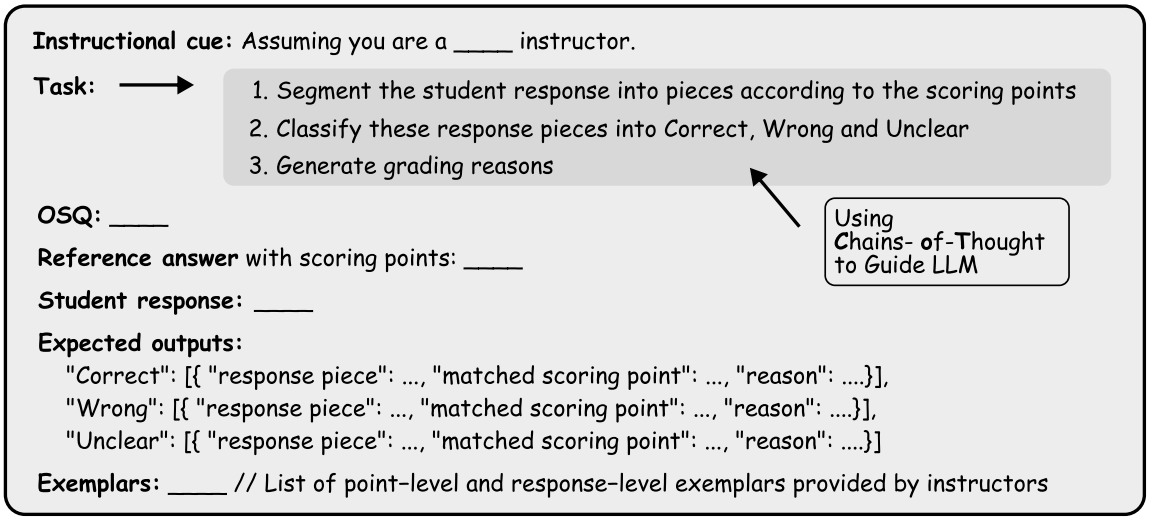


Fig. 3. Prompt design for automated grading, including instructional cue, OSQ, reference answers, student response, and exemplars. The LLM is guided to segment responses, classify pieces into Correct/Wrong/Unclear, and generate grading reasons using chain-of-thought reasoning.

grading logic. By integrating both, the prompt delivers layered guidance that ensures the LLM's outputs are logically sound, interpretable at the micro level, and standardized and coherent at the macro level.

### 5.4 User Interface

The *VeriGrader* interface is organized into two main views, namely, the navigation view (Figure 4-A) and grading view (Figure 4-B).

*5.4.1* ***Navigation View.*** This view serves as the central control hub of the *VeriGrader* system.

The navigation bar, positioned at the top of the interface, allows instructors to switch between questions using a tab-based selector and navigate student responses through a row of numbered circular buttons (Figure 4-A1). As shown in Figure 5, each response button is color-coded to indicate grading status: blue for the currently active response, green for responses preliminarily graded by the LLM, and yellow for exemplar responses verified by the instructor and designated as grading standards. Additionally, once the instructor confirms a response as fully graded, the corresponding button displays a blue outline. This visual encoding provides instructors with an immediate sense of grading progress.

Furthermore, to ensure that instructors can focus first on responses with potentially unreliable grading quality, we introduce *quality priority indicators*. These indicators do not evaluate semantic correctness but instead capture surface-level technical issues in the LLM outputs. Specifically, UR Rate measures the proportion of entities classified as unclear; high values indicate either that the LLM struggles to make decisive judgments or that the student response lacks clarity. Overlap Rate examines whether response pieces overlap within a student response; non-zero values suggest inconsistent boundary recognition and potential classification conflicts. Each student response is automatically assigned a priority level based on these metrics:

- High Priority (red dot): Overlap Rate $> 0$, or UR Rate $> 50\%$.
- Medium Priority (orange dot): UR Rate between 25–50%.

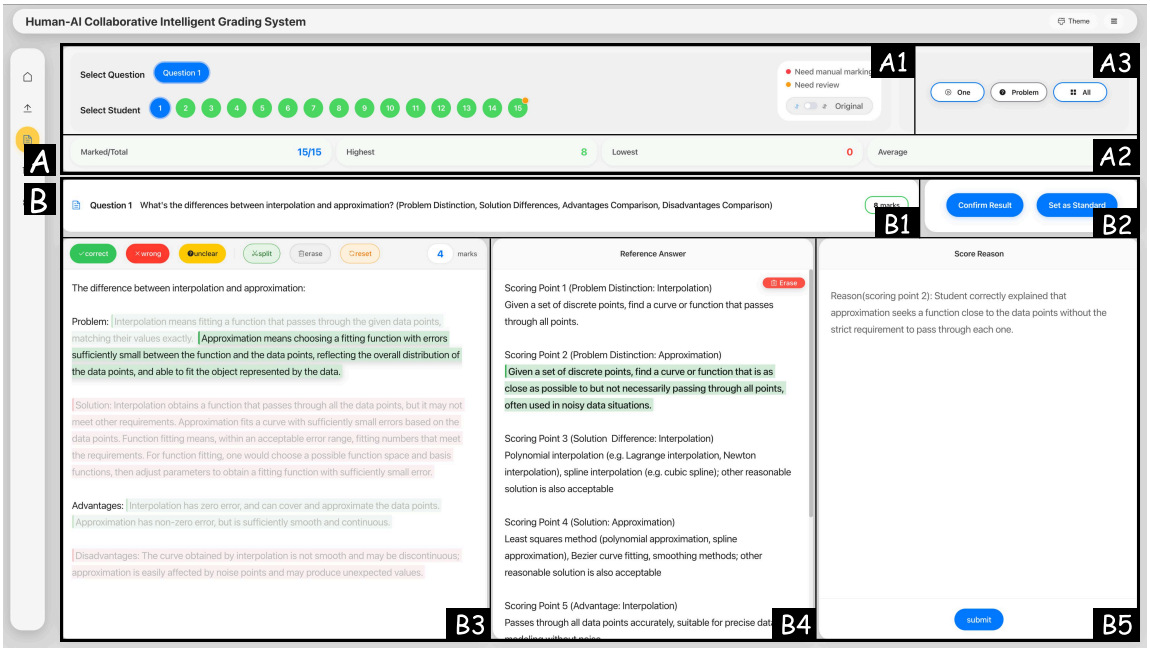


Fig. 4. *VeriGrader* System Interface. The interface comprises two views: a navigation view (A) and a grading view (B). In the navigation view, A1 provides a tab-based question selector and numbered response buttons, with a toggle for quality-based sorting; A2 summarizes performance with min/avg/max scores; A3 exposes three grading scopes: current , priority subset, and all non-confirmed responses. In the grading view, B1 shows the OSQ text; B2 lets instructors confirm a grading result or designate the current response as an exemplar. The response panel (B3) is the primary workspace, rendering LLM-extracted response pieces with category highlights. The answer panel (B4) presents the reference answer and highlights scoring points aligned to the selected response pieces, while the reason panel (B5) displays editable, LLM-generated grading reason for the current selection.


Fig. 5. Visual encoding and priority indicators

- Low Priority (no dot): all indicators fall within the normal range.

The toggle button in (Figure 4-A1) activates quality-based sorting, which reorders the responses such that exemplar responses appear first, followed by high-priority responses, then medium-priority responses, and finally low-priority responses. This mechanism directs instructor attention toward responses most likely to require review, ensuring efficient allocation of grading time.

Grading results for the current question are summarized with the highest, lowest, and average scores across all responses, thereby offering a concise overview of performance distribution (Figure 4-A2).

This view also provides useful core grading functions. It contains three categories of controls with three buttons (Figure 4-A3). First, clicking the “One” button triggers LLM’s regrading on the currently selected student response. Second, clicking the “Problem” button triggers LLM’s regrading on both high priority and medium priority responses. Third, clicking the “All” button triggers LLM to grade all responses except those that have already been confirmed,

including those designated as grading standard. These functions are particularly useful after the instructor has reviewed part of the LLM's grading results and provided exemplars for few-shot learning.

*5.4.2* ***Grading View.*** This view serves as the primary interface for the instructor to interact with the LLM and perform visual, interactive grading.

The selected question and its total score are displayed at the top of this view (Figure 4-B1), allowing the instructor to recall the question itself during the grading process. Next to them are two buttons (Figure 4-B2). If the instructor considers the grading result satisfactory, either produced directly by the LLM or refined through manual modification, the "confirm" button can be clicked to confirm it. The "Set as Standard" button allows the instructor to designate the current high-quality grading result as an exemplar response. Once clicked, the corresponding response button on the navigation bar simultaneously turns yellow. The response is then locked from further modification and incorporated into the grading template as a reference exemplar for the remainder of the grading process. This ensures both the authority of the exemplar and the consistency of LLM-assisted grading.

The lower part of this view is composed of three panels, namely, the response, answer, and reason panels.

**Response Panel.** The response panel (Figure 4-B3) serves as the primary panel of *VeriGrader*. The score of the student response is displayed in the menu of this panel and automatically updates in real-time based on the instructor's operations. This panel highlights the response pieces of each student response extracted by the LLM based on the scoring points. Response pieces are color-coded by category: correct (green), wrong (red), and unclear (yellow). When response pieces overlap with each other, the system orders them by start position and then by length, so that the shorter entity is displayed on top to maintain consistent visualization.

**Answer Panel.** This panel (Figure 4-B4) presents the reference answer for the current OSQ. It is infeasible for instructors to memorize the details of the reference answer. Even experienced instructors frequently consult it during the grading process. This panel also provides selection and highlighting features, making it easier for instructors to connect student responses to scoring points during the grading process.

**Reason Panel.** The reason panel (Figure 4-B5) is dedicated to displaying and managing the grading reason associated with the currently selected response piece. Specifically, the reason explains why the response piece is classified into a given category (correct, wrong, or unclear), providing transparency into the grading decisions. The reason can be edited by the instructor or requested to be generated by the LLM again, based on the response piece and scoring point.

**Cross-Panel Interactions.** We implement flexible interactions among the response, answer, and reason panels, enabling instructors to refine the LLM's grading results and, in turn, guide the model to regrade student responses. As shown in Figures 4-B3, B4, and B5, when an instructor hovers over a response piece in the response panel ("what was graded"), the corresponding scoring point in the answer panel is highlighted in the same style ("where is the scoring point") and the reason for this grading is also displayed in the reason panel ("why such a grading was made"). Based on this linkage, instructors can investigate the grading results and make several types of modifications as needed.

First, if a response piece is misclassified (e.g., marked as unclear), the instructor can reassign it to the correct category (e.g., correct) by selecting the appropriate label in the response panel toolbar. Upon reassignment, the system invokes the LLM to re-infer the alignment of the updated response piece with the relevant scoring points defined in the reference answers, and automatically regenerate the corresponding grading reason, while also allowing the instructor to refine or manually edit the explanation in the reason panel. Second, if a response piece is identified as irrelevant to the question, it can be deleted, with its highlight removed accordingly. Third, if a response piece is mapped to an incorrect scoring point, the instructor can remap it by selecting the response piece and brushing the appropriate scoring point in the

answer panel. The reasoning can then be updated to reflect the new mapping. Finally, if the LLM overlooks a text span that merits credit, the instructor can manually highlight the span, assign it the correct category, and generate the corresponding reasoning.

Regardless of the modification type, the three-panel linkage ensures that updates are propagated across the response, answer, and reason panels, maintaining consistency with the underlying data structure and preserving the completeness of scoring point coverage.

**Interactive User Feedback Incorporation.** As mentioned before, two kinds of exemplars can be incorporated into the LLM for better automated grading. As the user feedback is generated in this view, we implemented two interactions accordingly to support the incorporation of point-level and response-level exemplars. First, if the instructor identifies a quadruple <response piece, scoring point, category, reason> as representative and beneficial for enhancing the LLM's capability, it can be submitted via the "Submit" button in the reason panel, thereby adding it to the list of point-level exemplars. Second, if the entire student response is graded satisfactorily, the instructor may click the "Set as Standard" button (Figure 4-B2), which aggregates all quadruples in the current grading result into a response-level exemplar and adds it to the exemplar list. Based on the added exemplars, the instructor can click the buttons in the navigation view to trigger the LLM regrading procedure, obtaining better grading results.

Crucially, this process is incremental and continuous rather than one-off. As exemplars accumulate, the system gradually evolves from applying general grading to acting as a personalized, professional grading assistant. Over time, *VeriGrader* does not merely grade more accurately, but learns to grade like the instructor, ensuring outputs aligned with instructional intent.

### 5.5 Implementation

*VeriGrader* is implemented as a client-side web application following a serverless paradigm, directly interfacing with OpenAI-compatible LLM APIs for grading and pedagogical reasoning. The frontend, built with Vue 3 [40] and TypeScript [30], employs Pinia for state management and Element Plus for accessible UI components, with Vite serving as the build toolchain. A unified abstraction layer manages API orchestration, error handling, and response parsing, currently leveraging GPT-o3 models but extensible to alternative providers. Data persistence relies on browser native storage, eliminating backend infrastructure while ensuring privacy and low-latency operation.

## 6 Usage Scenario

To illustrate the effectiveness and usability of *VeriGrader*, we present a usage scenario in which we follow Daniel, a computer science professor, to see how he graded 30 student responses of an OSQ using *VeriGrader*. The OSQ is designed to assess students' understanding of interpolation and approximation algorithms given a set of discrete points. To ensure fairness and transparency, grading must consider all scoring points with explicit reasons. However, traditional manual grading is both time-consuming and exhausting. Thus, Daniel utilized the *VeriGrader* system to perform the grading process.

Daniel uploads the responses and reference answers to *VeriGrader*. Within 30 seconds, the LLM completes an initial round of grading. After Daniel issued the priority sorting, the system immediately reordered the responses, placing those with potential issues at the first. One response (#id = 15) was identified as medium priority (Figure 6) as the system identifies two response pieces as unclear (Figure 6-A1).

Daniel clicked on an unclear highlighted response piece, and the corresponding scoring point in the answer panel is simultaneously highlighted. Based on his expertise, Daniel determines that both *unclear* pieces are actually wrong,

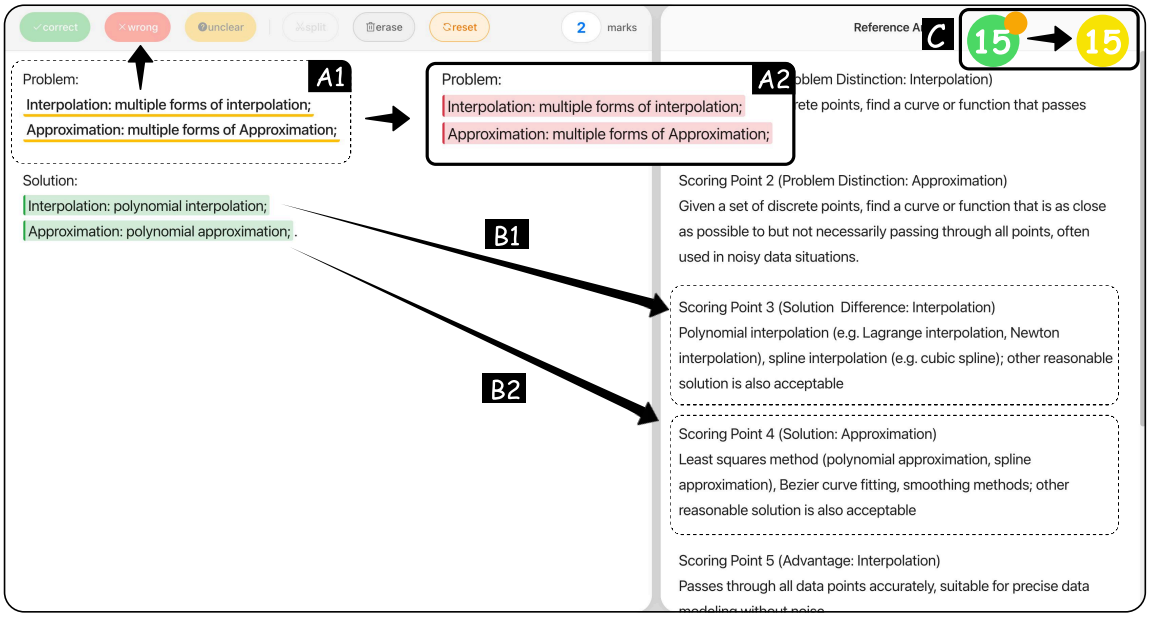


Fig. 6. Resolving unclear response pieces. Two initially unclear response pieces (A1) are manually reassigned to the *wrong* category (A2). The system retains the two remaining correct response pieces (B1–B2) with their associated reference answers. Finally, the instructor confirms the response as an exemplar (C), establishing it as a standard for subsequent grading.

as "*multiple forms of interpolation*" and "*multiple forms of Approximation*" are too vague to correspond to any valid scoring points. Then he relabels them as wrong (Figure 6-A2). The system automatically regenerates the associated grading reason. Afterwards, he checked the remaining two correct highlights and checked if they matched the correct scoring points. Once verified, Daniel considered these two were correct. In short, this response had both correct and wrong response pieces, and the wrong ones can not be determined by the LLM at the beginning. Daniel considered it a representative grading result that may help improve the LLM's performance and set it as a standard, establishing the first high-quality exemplar (Figure 6-C).

For the response of another student (#id = 1), Daniel again relied on the system's highlighting and reasoning support to quickly continue his assessment. When assessing the student response about the advantages of approximation, Daniel noticed that the LLM had not recognized "*Approximation has non-zero error, but is sufficiently smooth and continuous*" as a scoring point (Figure 7-A1). Upon analysis, he judged the answer to be correct and manually labeled it as correct by first brushing to select the text and then clicking the corresponding category button (Figure 7-A2). At that moment, *VeriGrader* automatically invokes the LLM to highlight the corresponding scoring point (Figure 7-A3) and generate a grading reason, while also updating the total score (Figure 7-C). Although the LLM can capture many correct mappings, there were still mismatches that required Daniel to verify and intervene to ensure reliability. After verifying that all highlighted response pieces accurately aligned with the corresponding scoring points (Figures 7-B1, B2, and B3), he determined that no further issues remained. He then designated this grading result as an exemplar (Figure 7-D), bringing the total to two high-quality exemplars within the system.

The third response presented a new case. The student mentioned "Overfitting (When noise is present)" (Figure 8-A) as a disadvantage of interpolation, while the reference answer specified "Sensitive to noise, high-order interpolation

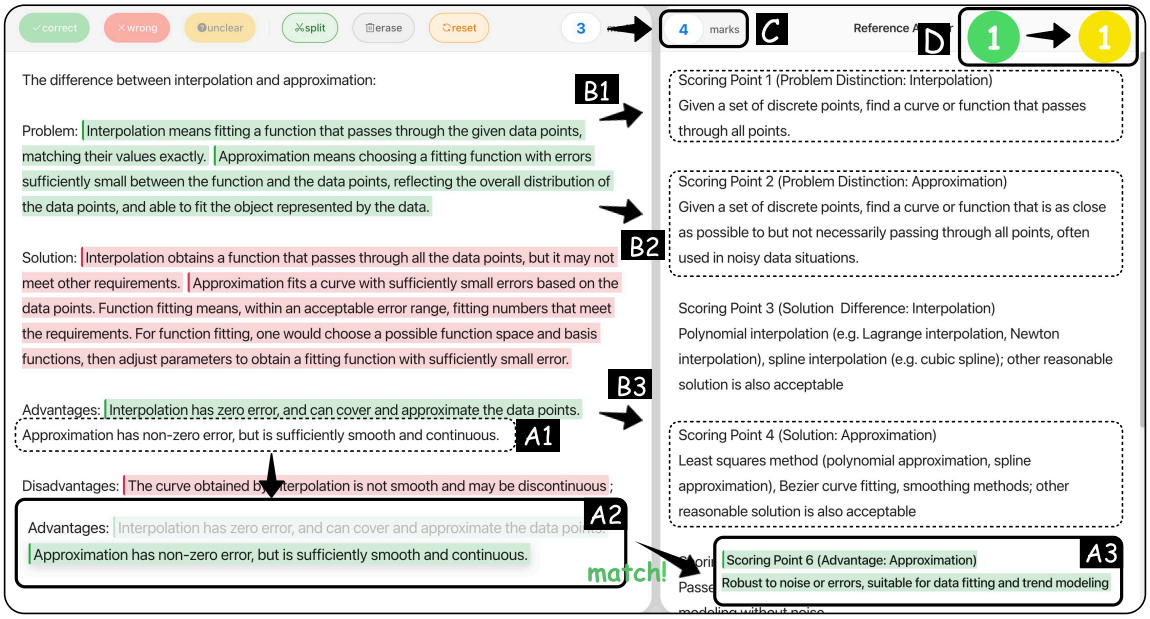


Fig. 7. Add the missing scoring points. An initially unscored response piece (A1) is manually marked correct and highlighted as A2, prompting the system to match the corresponding scoring point from the reference answer (A3) and update the total score (C). Together with the three previously recognized mappings (B1–B3), the newly added mapping (A2–A3) illustrates how C (the four points) are derived. With the scoring sources clearly established, the response is then designated as an exemplar (D).

may cause oscillations" as the scoring point (Figure 8-C). The system initially does not award credit, but Daniel judged that the two were essentially equivalent. The LLM failed to recognize such a semantic equivalence. So he brushed this response piece (Figure 8-A) to mark it as correct (Figure 8-B) and edited the reason panel to add: "Overfitting reflects sensitivity to noisy data, which is fundamentally the same as the scoring point 'sensitivity to noise' and should receive credit" (Figure 8-E). Daniel then submitted this customized reasoning to the system as a point-level exemplar, enabling *VeriGrader* to automatically recognize any similar response piece in the future. Finally, he confirmed the result and moved to the next response.

So far, the system has accumulated two response-level exemplars and one point-level exemplar. Daniel clicked the "all" button, requesting the system to regrade remaining responses with these exemplars for in-context learning. Daniel then adopted an iterative workflow: (1) he reviewed several responses, confirmed those that were accurate, (2) and when encountering special cases, added new exemplars before triggering another round of regrading. In this way, instructor-LLM collaboration progressively refines the entire grading process. After several iterations and accompanied by reviewing, Daniel completes grading all responses. Each response was associated with clear and explicit scoring points and reasons. *VeriGrader* not only improves efficiency but also internalizes Daniel's grading standards and pedagogical principles, thereby ensuring consistency and fairness in the final outcomes.

## 7 User Study

To evaluate the effectiveness of *VeriGrader*, we designed a controlled study. In this study, we compared *VeriGrader* grading of open-structured questions (OSQs) with a manual grading baseline system, which was developed to simulate

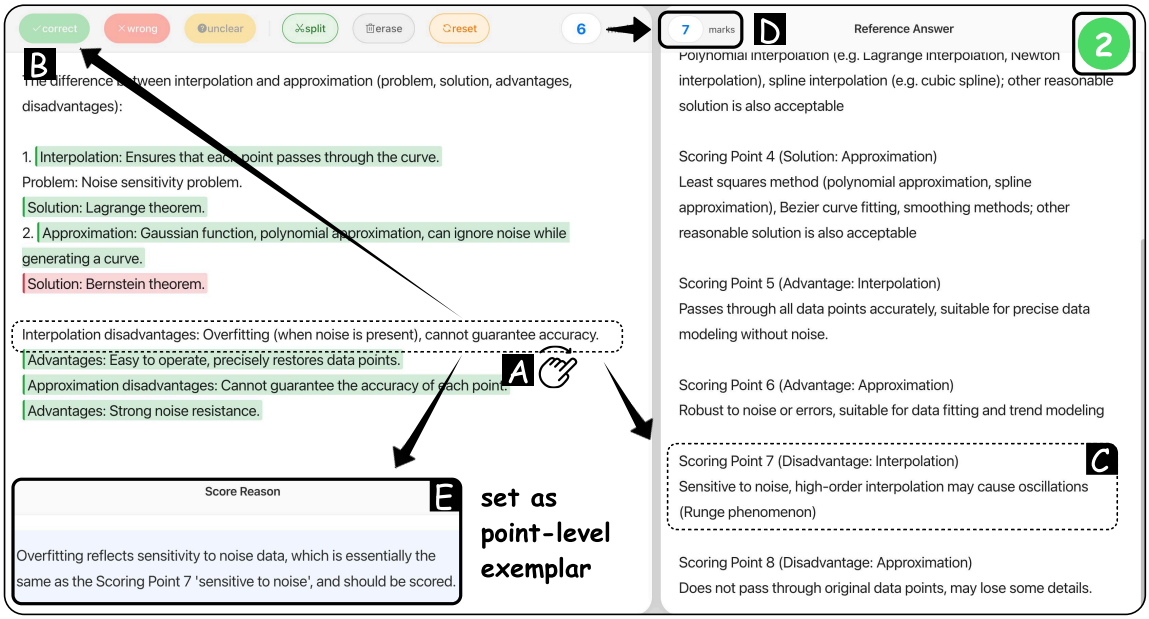


Fig. 8. The instructor highlights a student's response piece by dragging the mouse (A). He then assigns the appropriate category (B) , triggering the system to automatically invoke the model. As a result, a corresponding reference answer (C) and grading reason are generated, and the score is updated (D). To ensure consistency in future grading, the instructor further edits the rationale and saves it as a point-level exemplar (E).

the traditional manual grading process and ensure a fair comparison. Based on this setup, we formulated three research questions:

- RQ1: Can *VeriGrader* outperform manual grading in terms of accuracy and efficiency?
- RQ2: Does introducing few-shot exemplars lead to improvements in the model's grading performance?
- RQ3: Under absolutely no human intervention, can LLMs achieve great grading performance?

### 7.1 Study Setup

**Dataset.** We compiled a small-scale yet representative dataset that integrates carefully designed open-ended structured questions, instructor-authored reference answers with scoring points, and authentic student responses. The dataset contains two OSQs, selected for their coverage of distinct knowledge domains, namely **numerical methods** and **data structures**, and for their suitability for fine-grained scoring. Importantly, both sets of data were collected from actual in-class quizzes, reflecting real-world educational contexts. These responses capture diverse answering styles, such as partial correctness, misunderstanding of knowledge, and unclear or casual expressions, enabling a realistic evaluation of the system's ability to handle incomplete, noisy, or ambiguous inputs. All student participants were informed about the study, and their responses were collected with consent. To keep the experiment duration within a manageable range, we randomly selected 15 student responses for each question.

**Participants.** We recruited 6 participants (U1–U6; 4 males and 2 females), all of whom were teaching assistants with experience in grading. Participants are majoring in either software engineering or computer science, have experience

using interactive tools in their daily teaching or research, and have previously encountered or used AI-based tools, such as ChatGPT, DeepSeek, and Segment Anything.

Although the sample size is modest, recruiting teaching assistants with authentic prior grading experience was non-trivial in our context. We therefore prioritized depth over breadth, each participant completed a detailed four-step study, yielding a comprehensive evidence base.

**Systems and Grading settings.** To fully evaluate *VeriGrader*, we simplified the *VeriGrader*, removed its LLM-assisted functions, and replaced it with manual annotations, resulting in a manual grading system. Specifically, in the manual grading system, users only manually brush pieces of student responses in the response panel and label them as correct, wrong, or unclear, while brushing to select the corresponding scoring points from the reference answers in the answer panel. Based on these two systems, we designed three grading settings: (**S1**) manual grading, (**S2**) *VeriGrader* grading in a normal usage process, i.e., without high-quality few-shot exemplars at initialization, and (**S3**) *VeriGrader* grading with high-quality few-shot exemplars provided in advance.

To compare *VeriGrader* and the manual grading system (**RQ1**), each participant should grade responses under Setting **S1** and under Setting **S2**, respectively. To examine the effectiveness of incorporating user feedback (**RQ2**), each participant graded ten responses under Setting **S3**. Particularly, the few-shot exemplars in Setting **S3** should be generated by the participants themselves under Setting **S2**. To mitigate learning effects, the question alternated between data structures and numerical methods across different participants and settings. In addition, the order in which systems were used was counterbalanced among participants. Given the considerations above, the two experiments were integrated into a unified experimental procedure. After the experiments, we performed post-quantitative analysis of the LLM's grading results to address (**RQ3**).

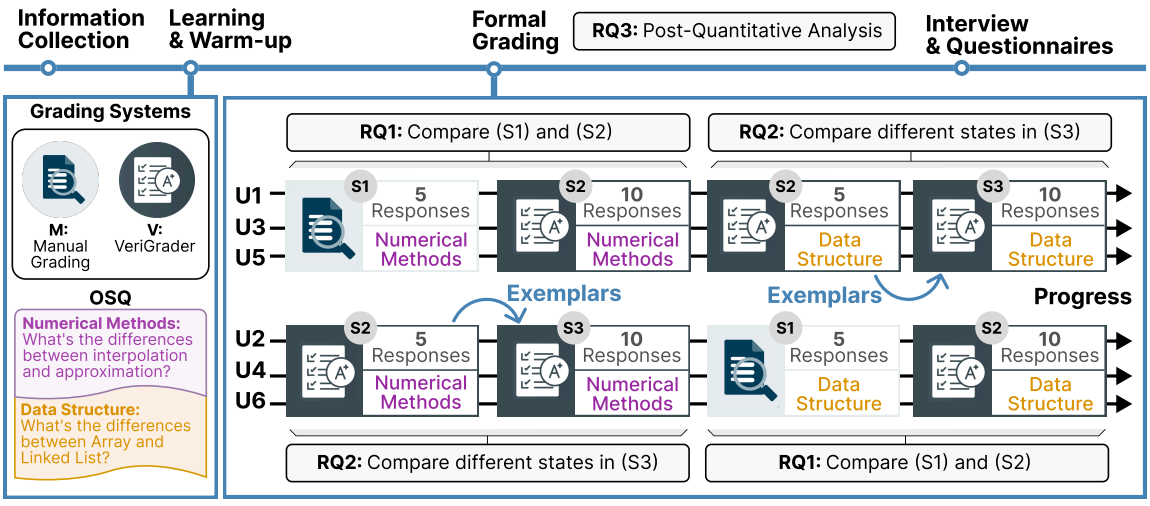


Fig. 9. The experiment procedure consisted of four sequential steps: (1) Information Collection, obtaining consent and gathering participant background; (2) Learning and Warm-up, introducing system functions and practicing with sample OSQ; (3) Formal Grading, performing detailed grading of student responses under both manual and *VeriGrader* settings; and (4) Interviews and Questionnaires, collecting qualitative feedback and usability evaluations.

**Procedure.** The study followed a structured four-step protocol (Figure 9).

**Step 1: Information Collection** (5 minutes). First, participants provided informed consent and were briefed on the study objectives, tasks, and potential risks. Their personal information, like major and familiarity with AI-based tools, was also collected.

**Step 2: Learning and Warm-up** (30 minutes) Next, we introduced the core functions, visualization, and interactions of the two grading systems. During this step, we provide an additional OSQ for warm-up for both grading modes to ensure that participants are proficient in the functions, visualizations, and interactions. Finally, participants were required to fully understand the two questions they would be grading and the corresponding reference answers. This learning and warm-up step is carried out in detail to minimize subsequent learning effects.

**Step 3: Formal Grading** (75 minutes). Third, we provided participants with two OSQs, for each of which 15 student responses were collected. Participants were instructed to perform fine-grained and explainable grading of each student response. They were asked to align as much content as possible from the students' answers with the predefined scoring points and to evaluate them accurately. When a sentence contained multiple scoring points, each point was assessed individually as correct, wrong, or unclear. This procedure ensured that all scoring points were consistently and precisely evaluated. In order to answer the three research questions simultaneously and reduce the learning effect, we designed the experimental pipeline shown in Figure 9. Each user goes through both the manual grading system (Setting **S1**) and *VeriGrader* (Setting **S2**). The order of the systems varies for different users. Particularly, the grading results generated by the user using *VeriGrader* in Setting **S2** will be immediately used for the process in which the user uses *VeriGrader* in Setting **S3**. As a result, each participant was required to complete the grading of four batches of student responses.

**Step 4: Interviews and Questionnaires** (10 minutes) Finally, we interviewed each participant to collect their feedback on both grading modes. Each participant completed a tailored Likert-scale questionnaire and two separate System Usability Scale (SUS) questionnaires: one for the manual grading system and the other for *VeriGrader*. The entire experiment lasted approximately 120 minutes, and each participant received a compensation of $20.

### 7.2 Measurement

**Ground Truth.** To establish a reliable evaluation benchmark, we invited a domain expert with extensive teaching experience to conduct fine-grained manual grading on all student responses. Following authentic grading practices, the expert carefully compared each student response $r_i$ against the reference answers and identified which scoring points were covered. Consequently, for each response, we had $\text{GroundTruth}_i$ also in a form of the tuple <response piece, scoring point, category>.

**Accuracy.** For each student response $r_i$, we model each participant's grading result as $\text{Result}_i$ in the form of the tuple <response piece, scoring point, category>. To comprehensively evaluate the quality of identified scoring points, we compared the system or manual grading results against the ground truth. Let $n$ denote the total number of student responses in a batch. *F1-score* was used as the measure of grading accuracy as follows:

$$\text{Precision} = \frac{\left|\bigcup_{i=1}^{n}(\text{Result}_i \cap \text{GroundTruth}_i)\right|}{\left|\bigcup_{i=1}^{n}\text{Result}_i\right|} \tag{1}$$

$$\text{Recall} = \frac{\left|\bigcup_{i=1}^{n}(\text{Result}_i \cap \text{GroundTruth}_i)\right|}{\left|\bigcup_{i=1}^{n}\text{GroundTruth}_i\right|} \tag{2}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Particularly, when computing intersections, we need to determine whether two text segments, such as a response piece from the ground truth and one from the system output, are consistent. We first generate preliminary matches using string matching, and then perform a manual verification based on these initial results.

**Efficiency.** For each response, the time spent was measured from the moment the user first views the question to the moment they move on to the next one. We used the *average time spent per OSQ* (minutes/OSQ) as the efficiency measure.

### 7.3 Quantitative Result

The quantitative analysis of the user study results addresses the research questions mentioned before.

**RQ1: *VeriGrader* outperformed the manual baseline in both efficiency and accuracy.** We compared the grading results under Setting **S1** and Setting **S2**. The latter reflects the common human–AI collaborative workflow in *VeriGrader*, while the former serves as the manual grading baseline. The results (Figure 10) indicated that *VeriGrader* outperformed the manual system in both efficiency and accuracy, and these advantages can be achieved simultaneously. *VeriGrader* generally reduced grading time without sacrificing accuracy, and in most cases, it even enhanced grading quality.

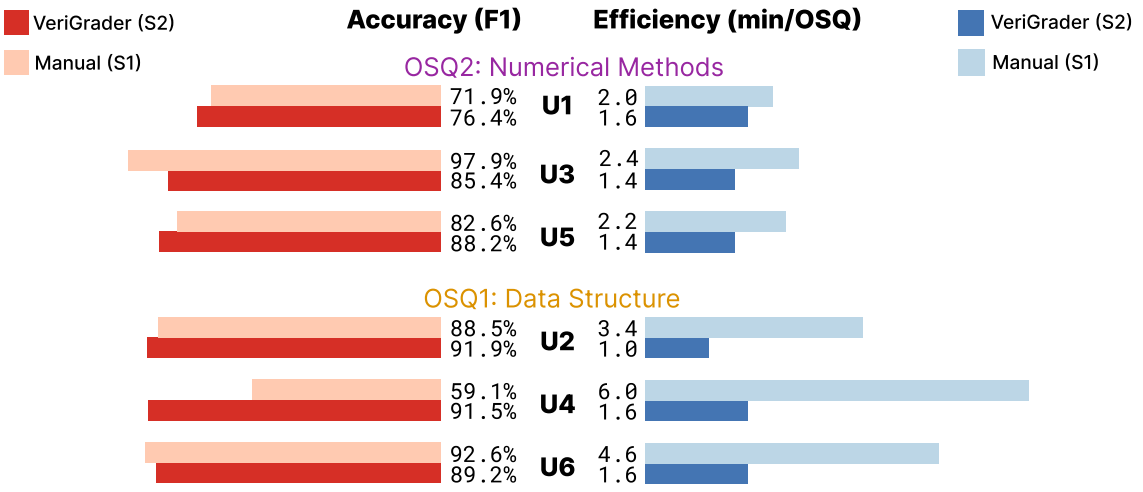


Fig. 10. Comparison of Setting **S1** and Setting **S2** on OSQ1 (data structure) and OSQ2 (numerical methods). Across all users, *VeriGrader* substantially reduced grading time (e.g., U4 from 6.0 to 1.6 minutes, a 73% reduction) while maintaining or improving accuracy. Although minor decreases occurred in a few cases (e.g., U3 and U6), the overall trend demonstrates that *VeriGrader* simultaneously improves efficiency and accuracy compared to the manual Setting **S1**.

In terms of efficiency, *VeriGrader* obviously reduced grading time, with decreases ranging from 30% to 70% in most cases. As shown in Figure 10, for OSQ1 (data structure), U4's average grading time dropped from 6.0 minutes to 1.6 minutes (-73% overall), U2 from 3.4 minutes to 1.0 minute (-71% overall), and U6 from 4.6 minutes to 1.6 minutes (-65% overall). In OSQ2 (numerical methods), U3's time decreased from 2.4 to 1.4 minutes (-42% overall), U1 from 2.0 to 1.6 minutes (-20% overall), and U5 from 2.2 to 1.4 minutes (-36% overall). These results show that while instructors exhibited large variations in manual grading times (ranging from 2 to 6 minutes per response), the use of *VeriGrader* stabilized performance across participants, with all users converging to approximately 1–1.6 minutes per response.

In terms of accuracy, *VeriGrader* generally outperformed manual grading For OSQ1, U4's accuracy increased markedly from 59.1% to 91.5%, and U2 improved from 88.5% to 91.9%. For OSQ2, U1 rose from 71.9% to 76.4%, and U5 from 82.6% to 88.2%. Although a few participants exhibited fluctuations (e.g., U3 from 97.9% to 85.4%, U6 from 92.6% to 89.2%), the overall average accuracy improved to 87.1%, higher than the manual baseline. Moreover, the variance across instructors was reduced, as accuracy scores converged after using *VeriGrader*, suggesting that the system not only elevates overall performance but also promotes consistency and fairness in grading outcomes.

**RQ2: Introducing few-shot exemplars consistently improved grading accuracy.** To verify the incorporation of user feedback, we recorded the grading results under three conditions with 10 student responses for each participant under Setting **S3**, including (1) the initial grading results from the LLM without any exemplars, (2) the grading results from the LLM with exemplars generated by the participants under Setting **S2**, and (3) the final grading results confirmed by the participants. Figure 11 presents the accuracy of these results. Overall, few-shot exemplars consistently improved accuracy, with the effect being more pronounced in the more complex OSQ1 (data structure). Moreover, when combined with final human review, accuracy gains were further reinforced in most cases, suggesting that human–AI collaboration can enhance grading reliability beyond model-only improvements.

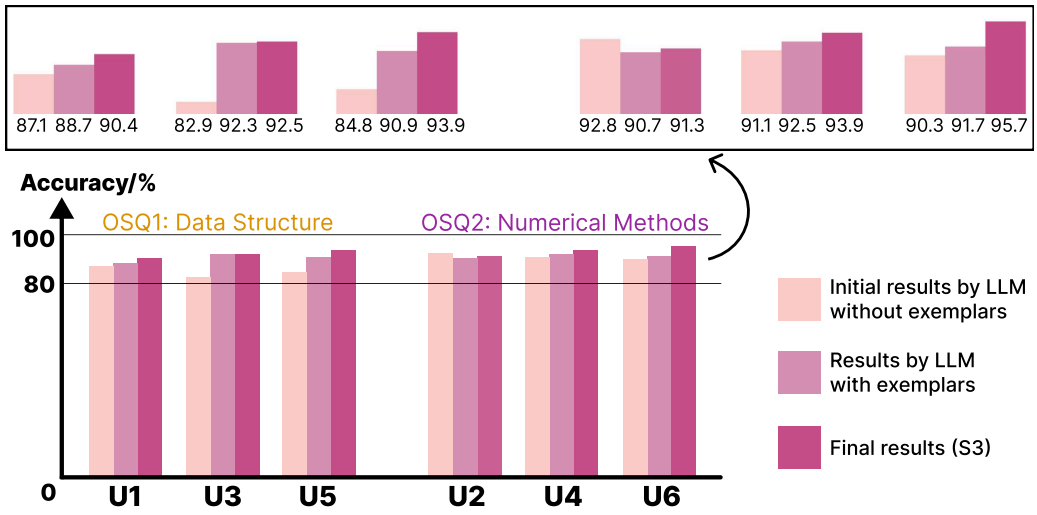


Fig. 11. Grading accuracy with and without few-shot exemplars, and with additional human review and refinement, across OSQ1 (data structure) and OSQ2 (numerical methods). The introduction of few-shot exemplars consistently improved accuracy across all participants, with more pronounced gains in OSQ1, where performance increases reached nearly 10% (U3). In contrast, improvements in OSQ2 were more modest due to its higher baseline, though human review still elevated accuracy to above 95% for selected participants (U6). These findings demonstrate that few-shot exemplars are particularly effective for complex, lower-baseline tasks, while review and refinement provide an additional layer of reliability in both task types.

In terms of distribution, three participants who reviewed OSQ1 achieved average accuracy improvements of more than 5% after introducing few-shot exemplars. U1 improved from 87.1% to 88.7% and further to 90.4% with review (+3.3% overall). U3 exhibited a substantial gain after introducing few-shot exemplars (+ 9.4% overall), and further approached 93% accuracy after the review stage, while U5 showed a similar trajectory with steady increases across all conditions, from 84.8% to 90.9% and 93.9% (+9.1% overall). This pattern highlights the particular benefits of exemplars when handling longer and more complex responses. By contrast, the baseline accuracy for OSQ2 was already high (average 89.5%), so

the marginal effect of introducing few-shot exemplars was smaller but remained consistently positive. U2 vibrated at a relatively high level (average 91.6%), with an error of about 1%. We considered this a reasonable phenomenon. Still, U4 improved from 91.1% to 92.5% and 93.9% (+2.8% overall), and U6 advanced from 90.3% to 91.7% after introducing few-shot learning. Notably, U6 still achieved further improvements with the review and on-demand refinement, surpassing 95% accuracy, showing that exemplars and human intervention can yield further improvements even in high-baseline settings.

Taken together, these findings indicate that the incorporation of user feedback not only mitigated model limitations in more demanding tasks, but also sustain reliable gains in high-baseline contexts, thereby supporting a robust framework for human–AI collaborative grading.

**RQ3: Preliminary Effectiveness of the LLM itself**. We evaluated the LLM using the second and fourth batches of student responses, consisting of 10 responses for OSQ1 and 10 responses for OSQ2. Six rounds of initial LLM grading results were collected, each corresponding to a participant's grading session. We calculated the average grading accuracy across the six rounds for each OSQ to evaluate whether the LLM can perform effective grading without human intervention, while reducing noise and mitigating uncertainty in the model's outputs. The LLM achieved an average accuracy of 87.5% on OSQ1 (data structure) and 89.6% on OSQ2 (numerical methods). These results demonstrate that the model maintains high accuracy even in a fully automated setting, indicating its potential practical value. Nonetheless, as discussed earlier, human involvement can further improve grading outcomes, which is particularly important in high-stakes scenarios.

### 7.4 User's Feedback

*VeriGrader* also received positive feedback from all participants. The detailed results of the SUS and tailored Likert-scale questionnaire are illustrated in Figure 12 and Figure 13.

**F1: Showing a "cautious trust" in the LLM capabilities.** All participants (6/6) acknowledged the understanding and classification capabilities of LLM in grading (Tailored Q1=6.17). However, their views on reliability varied (Tailored Q2=4.50). Some participants expressed relatively higher trust, while others emphasized the need for human supervision. The less trusting participants reported that they would review a subset of responses to determine whether the LLM's performance met their expectations. For example, U3 pointed out: "*The LLM considers multiple scoring points in a long sentence as only one single scoring point.*" Similarly, U4 stated: "*I would prioritize checking the places where there was no highlight, as they might contain missed scoring points.*" This attitude reflected their "cautious trust" in the LLM. Participants regarded the LLM as a supplementary aid rather than a fully reliable solution, aiming to ensure fairness and rigor in grading.

**F2: Using LLM to extract response pieces can improve efficiency.** *VeriGrader*'s automatic segmentation function for student response received positive feedback (4/6). Participants stated that this function largely reduced the burden of identifying scoring points in long answers one by one, enabling them to quickly locate the points extracted by the LLM, and then review them and make moderate adjustments. In addition, U2 emphasized that *VeriGrader*'s automatic deduplication mechanism effectively alleviated the tediousness of duplicate scoring: "*In manual scoring, I must tediously identify pieces that correspond to the same scoring point and assign credit only once, while the system can complete this process automatically.*" The tailored questionnaire results (Figure 13) confirmed this advantage: Tailored Q3=6.00 (efficiency improvement) and Tailored Q4=6.33 (more accurate capture of scoring points).

**F3: Intuitive visualization and interactive experience.** The mapping between response pieces and highlighted reference answers was regarded as intuitive and traceable, helping participants quickly locate the source of scores. For

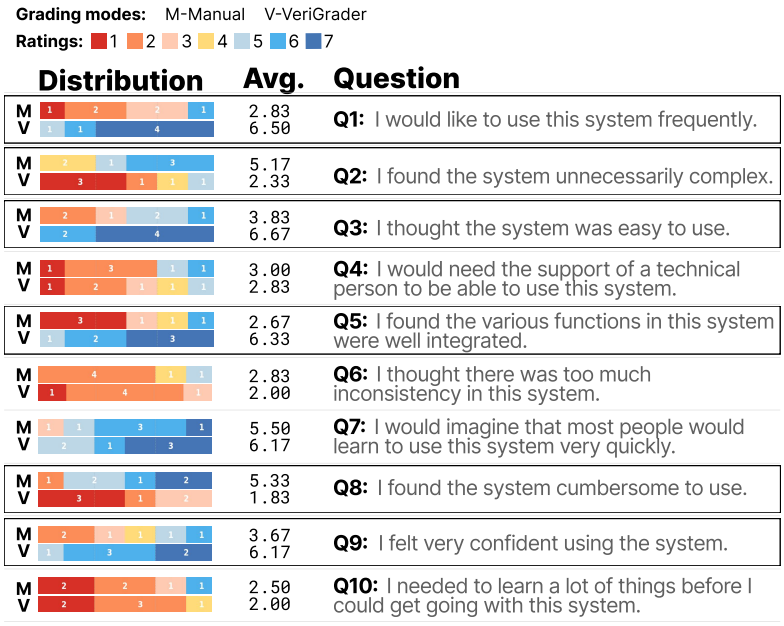


Fig. 12. Results of the SUS questionnaire. The boxes indicate cases where the average user ratings of the two systems differ by more than 2 points and the Wilcoxon test shows a significant difference ($p<0.05$).

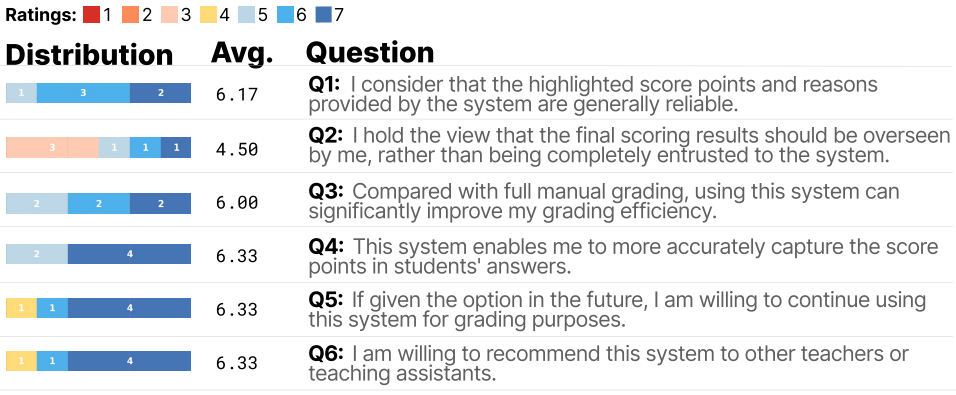


Fig. 13. Participants' responses to the 7-point Likert-scale questionnaire items assessing. Higher scores indicate stronger agreement.

example, U1 said: "*I clearly knew where the score came from.*" The SUS results (Figure 13) also supported this finding. *VeriGrader* achieved much higher ratings than the manual system on usability (Q3=6.67 vs. 3.83) and function integration (Q5=6.33 vs. 2.67). For willingness to use frequently, the manual system received a low score of 2.83, whereas *VeriGrader* reached 6.50.

**F4: Practical Utility and Adoption Intention.** Participants consistently expressed a strong willingness to adopt the system in their future grading practice. The tailored questionnaire results (Figure 13) showed high ratings for both

willingness to continue using the system (Q5, M=6.33/7) and willingness to recommend it to others (Q6, M=6.33/7). This positive adoption intention was primarily attributed to the system's practical utility: participants emphasized that it alleviated the burden of repetitive grading, improved efficiency (Q3, M=6.00), and enabled more accurate capture of scoring points (Q4, M=6.33). These findings highlight that participants valued the system as a useful and trustworthy tool within their workflow, even while maintaining a supervisory role as noted in T1.

Beyond practical utility, participants also highlighted the interactive and evolving nature of the system as an important factor in sustaining engagement. U3 said: "*I was constantly interacting, this was very fun.*" U5 further stated: "*I felt like I was teaching the model how to grade. By continuously adding high-quality exemplars, my own subjective trust in the model's ability to perform better also increased.*" These reflections suggest that the combination of practical benefits and interactive learning experiences shaped a positive attitude toward long-term adoption, blending utilitarian value with a sense of co-agency in grading.

## 8 Discussion

Assessing open-ended structured questions (OSQs) has long posed a complex challenge in education, requiring instructors to balance fair and accurate grading while accommodating the diversity of student responses. Through the design, implementation, and evaluation of *VeriGrader*, we gained deep insights into how human-AI collaboration operates in such high-stakes assessment scenarios. The following sections discuss this study from multiple perspectives.

### 8.1 Design Principles of Human-AI Collaboration

Our research reveals several key design principles for human-AI collaboration in educational assessment, a domain where fairness and accuracy are paramount.

First, a **transparency-first** principle is essential. Unlike "black-box" automated grading approaches, *VeriGrader* makes LLM decision-making fully visible through fine-grained response segmentation and scoring point mapping. Instructors retain ultimate authority for modifications and final decisions. This design not only meets the accountability requirements inherent in educational contexts but also provides instructors with clear evidence for evaluating and validating AI-assisted judgments.

Second, it is important to **foster user trust in AI from the outset**. We observed that instructors initially interacted with the system cautiously, but as the system progressively adapted to their grading preferences and expertise, both grading quality and user trust improved. This design mitigates the risks of full automation while enabling the system to evolve from a generic AI assistant into a personalized professional tool through iterative few-shot learning.

Third, **bidirectional feedback mechanisms** support effective collaboration. *VeriGrader* enables continuous interaction between instructors and LLMs. Instructor corrections not only fix immediate errors but also improve future decisions by enriching the in-context learning examples, while LLM outputs provide a clear reference for instructors to ensure grading consistency.

Finally, when extending AI-assisted systems to new application domains, maintaining **human professional authority** while leveraging AI for efficiency and consistency is essential. This collaborative paradigm is applicable across fields that require expert judgment with low tolerance for errors, such as medical diagnosis, legal document review, and financial risk assessment.

### 8.2 Technical Insights

From a technical perspective, the main challenge in OSQ assessment is balancing structured scoring with the open-ended nature of student responses. Our segmentation–mapping–classification approach addresses this challenge by breaking student responses into pieces that align with predefined scoring points, while still capturing diverse ways of expression.

Notably, we introduced an unclear category to fill the gap left by the traditional binary classification of correct and wrong. We refer to this design as "deliberate openness," emphasizing that the system purposefully leaves room for cases where student responses cannot be cleanly judged. Our analysis shows that this category not only captures ambiguity in student responses but also highlights opportunities for instructional intervention, as such responses often indicate subtle misunderstandings that require targeted guidance. In addition, each scoring decision is paired with explicit reasoning, which improves system transparency and gives instructors concrete targets for refining their feedback.

### 8.3 Limitations and Future Work

*VeriGrader* performed well and was well-received by users, but some limitations remain.

First, the system's grading effectiveness is closely tied to the structural clarity of student responses. When answers are well-organized with clear bullet points, the system can accurately identify and map scoring points; however, when responses are presented as lengthy texts containing multiple potential scoring points, the system's identification and matching quality declines. This observation aligns with human manual grading behavior, as instructors also tend to favor well-structured and clearly expressed answers over scattered or incomplete responses.

Second, the system currently applies primarily to OSQs with relatively explicit scoring points, with limited applicability to completely open-ended creative writing or critical thinking questions. This limitation stems mainly from such questions lacking standardized scoring references, making it difficult to establish stable segmentation-mapping frameworks. Notably, recent studies, such as CoGrader [5], have begun to explore grading for more open-ended responses, suggesting that future work could integrate these approaches and features to develop a unified grading system capable of handling a wider range of question types.

Third, beyond the system itself, even the human-review process revealed certain limitations when compared against the ground truth. Specifically, although instructors carefully reviewed the responses in the final stage, their judgments still did not achieve perfect alignment with the reference. This doesn't mean that human judgment has flaws, but rather reflects two underlying factors. First, in authentic grading scenarios, instructors often review the responses multiple times over an extended period before reaching a stable and consistent decision, whereas our experimental setting required completion within a constrained timeframe. Second, the ground truth was constructed with strict adherence to predefined scoring points, while instructors typically apply their own interpretations and provide reasonable justifications that may not fully match the reference. Consequently, under strict ground-truth comparison, these responses were marked as "incorrect", even though they remained pedagogically sound. This finding highlights that the definition of ground truth itself influences evaluation outcomes, suggesting that future work should explore multi-dimensional and tolerance-aware evaluation metrics.

Finally, the current system is primarily designed for text-based OSQs, offering limited support for responses that include diagrams, formulas, or other multimodal elements. Integrating multimodal assessment capabilities represents an important direction for future research. By incorporating visual understanding models, the system could process student responses containing hand-drawn illustrations, diagrams, or mathematical formulas, enabling multimodal assessment and feedback as well as more intuitive visualization and interaction. Such extensions would significantly

expand the system's applicability, particularly in science, technology, engineering, and mathematics (STEM) education, where non-textual responses are common.

### 8.4 Implication for Educational Technology Ecosystems

From a broader perspective, *VeriGrader* illustrates a key paradigm in educational technology: enhancing rather than replacing human expertise. Effective AI systems should form partnerships with educators rather than fully automate teaching processes. This collaborative approach suggests that future educators will need AI literacy and the ability to guide AI assistants to align with pedagogical goals, while institutions implement quality assurance mechanisms to maintain transparency and accountability. In the long term, educational ecosystems may comprise multiple specialized AI agents—such as grading, feedback, and learning path planning agents—working under human oversight to support personalized, high-quality learning. *VeriGrader* provides technical foundations and design insights relevant to developing such systems.

## 9 Conclusion

This paper presents the design, implementation, and evaluation of *VeriGrader*, an instructor–LLM collaborative grading system that addresses the challenges of assessing open-ended structured questions (OSQs). To ground our design, we first conducted preliminary interviews with stakeholders, along with the analysis of 141 graded exam papers, to deeply understand the actual workflow, core challenges, and potential improvement needs in grading OSQs, and identified instructors' expectations and concerns regarding AI-assisted grading tools. Based on these findings and requirement analysis, we designed and implemented *VeriGrader*. *VeriGrader* leverages the capability of LLMs to segment student responses to OSQs into pieces, align them with the scoring points of the reference answer, and ultimately score them with explanations. Furthermore, it enables instructors to review and refine LLM grading results through intuitive visual interfaces and incorporate instructor feedback into subsequent automated grading through in-context learning. We present a usage scenario to demonstrate *VeriGrader* and conduct a comprehensive user study with six teaching assistants to confirm the system's effectiveness. We also discussed observations and findings from the research process, which provide valuable insight into AI plus Education.

## References


[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–13. doi:10.1145/3290605.3300233

[2] Sami Baral, Eamon Worden, Wen-Chiang Lim, Zhuang Luo, Christopher Santorelli, and Ashish Gurung. 2024. Automated Assessment in Math Education: A Comparative Analysis of LLMs for Open-Ended Responses. In *Proceedings of the International Conference on Educational Data Mining*. International Educational Data Mining Society.

[3] Susan M Brookhart. 2010. *How to assess higher-order thinking skills in your classroom.* Ascd.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[5] Zixin Chen, Jiachen Wang, Yumeng Li, Haobo Li, Chuhan Shi, Rong Zhang, and Huamin Qu. 2025. CoGrader: Transforming Instructors' Assessment of Project Reports through Collaborative LLM Integration. *Proceedings of ACM Symposium on User Interface Software and Technology*.

[6] Clayton Cohn, Nicole Hutchins, Tuan Le, and Gautam Biswas. 2024. A Chain-of-Thought Prompting Approach with LLMs for Evaluating Students' Formative Assessment Responses in Science. In *AAAI Conference on Artificial Intelligence*. AAAI Press, 23182–23190. doi:10.1609/AAAI.V38I21.30364

[7] Linda Darling-Hammond. 2017. Developing and Measuring Higher Order Skills: Models for State Performance Assessment Systems. *Council of Chief State School Officers* (2017).

[8] Ehsan Doostmohammadi, Oskar Holmström, and Marco Kuhlmann. 2024. How Reliable Are Automatic Evaluation Methods for Instruction-Tuned LLMs?. In *Findings of the Association for Computational Linguistics: EMNLP*. Association for Computational Linguistics, 6321–6336. doi:10.18653/V1/

2024.FINDINGS-EMNLP.367

[9] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (2018), 1–37.

[10] Haoxiang Fan, Guanzheng Chen, Xingbo Wang, and Zhenhui Peng. 2024. LessonPlanner: Assisting Novice Teachers to Prepare Pedagogy-Driven Lesson Plans with Large Language Models. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*. ACM, 146:1–146:20. doi:10.1145/3654777.3676390

[11] Graham Gibbs and Claire Simpson. 2005. Conditions under which assessment supports students' learning. *Learning and teaching in higher education* 1 (2005), 3–31.

[12] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A Survey on LLM-as-a-Judge. *CoRR* abs/2411.15594 (2024). doi:10.48550/ARXIV.2411.15594

[13] Arto Hellas, Juho Leinonen, Sami Sarsa, Charles Koutcheme, Lilja Kujanpää, and Juha Sorva. 2023. Exploring the Responses of Large Language Models to Beginner Programmers' Help Requests. In *Proceedings of the 2023 ACM Conference on International Computing Education Research - Volume 1*, Kathi Fisler, Paul Denny, Diana Franklin, and Margaret Hamilton (Eds.). ACM, 93–105. doi:10.1145/3568813.3600139

[14] Owen Henkel, Libby Hills, Adam Boxer, Bill Roberts, and Zachary Levonian. 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 ACM Conference on Learning @ Scale,*. ACM, 300–304. doi:10.1145/3657604.3664693

[15] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain informatics* 3, 2 (2016), 119–131.

[16] Hung-Yu Huang. 2023. Modeling rating order effects under item response theory models for rater-mediated assessments. *Applied Psychological Measurement* 47, 4 (2023), 312–327.

[17] Yanwei Huang, Yunfan Zhou, Ran Chen, Changhao Pan, Xinhuan Shu, Di Weng, and Yingcai Wu. 2024. Interactive Table Synthesis With Natural Language. *IEEE Trans. Vis. Comput. Graph.* 30, 9 (2024), 6130–6145. doi:10.1109/TVCG.2023.3329120

[18] Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. CodeAid: Evaluating a Classroom Deployment of an LLM-based Programming Assistant that Balances Student and Educator Needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 650:1–650:20. doi:10.1145/3613904.3642773

[19] Jenia Kim, Henry Maathuis, and Danielle Sent. 2024. Human-centered evaluation of explainable AI applications: a systematic review. *Frontiers in Artificial Intelligence* 7 (2024), 1456486.

[20] Charles Koutcheme, Nicola Dainese, Sami Sarsa, Arto Hellas, Juho Leinonen, and Paul Denny. 2024. Open Source Language Models Can Provide Feedback: Evaluating LLMs' Ability to Help Students Using GPT-4-As-A-Judge. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*. ACM. doi:10.1145/3649217.3653612

[21] Jung X. Lee and Yeong-Tae Song. 2024. College Exam Grader using LLM AI models. In *2024 IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*. 282–289. doi:10.1109/SNPD61259.2024.10673924

[22] Pei Yee Liew and Ian K. T. Tan. 2024. On Automated Essay Grading using Large Language Models. In *Proceedings of the International Conference on Computer Science and Artificial Intelligence*. ACM, 204–211. doi:10.1145/3709026.3709030

[23] Guangming Ling, Pamela Mollaun, and Xiaoming Xi. 2014. A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing* 31, 4 (2014), 479–499.

[24] Ming Liu, Yiling Ren, Lucy Michael Nyagoga, Francis Stonier, Zhongming Wu, and Liang Yu. 2023. Future of education in the era of generative artificial intelligence: Consensus among Chinese scholars on applications of ChatGPT in schools. *Future in Educational Research* 1, 1 (2023), 72–101.

[25] Rongxin Liu, Carter Zenke, Charlie Liu, Andrew Holmes, Patrick Thornton, and David J. Malan. 2024. Teaching CS50 with AI: Leveraging Generative Artificial Intelligence in Computer Science Education. In *Proceedings of the ACM Technical Symposium on Computer Science Education*. ACM, 750–756. doi:10.1145/3626252.3630938

[26] Ziao Liu, Xiao Xie, Moqi He, Wenshuo Zhao, Yihong Wu, Liqi Cheng, Hui Zhang, and Yingcai Wu. 2025. Smartboard: Visual Exploration of Team Tactics with LLM Agent. *IEEE Trans. Vis. Comput. Graph.* 31, 1 (2025), 23–33. doi:10.1109/TVCG.2024.3456200

[27] Xinyi Lu, Simin Fan, Jessica Houghton, Lu Wang, and Xu Wang. 2023. ReadingQuizMaker: A Human-NLP Collaborative System that Supports Instructors to Design High-Quality Reading Quiz Questions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 454:1–454:18. doi:10.1145/3544548.3580957

[28] Kelly McConvey, Shion Guha, and Anastasia Kuzminykh. 2023. A Human-Centered Review of Algorithms in Decision-Making in Higher Education. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, 223:1–223:15. doi:10.1145/3544548.3580658

[29] Rafael Ferreira Mello, Cleon Pereira Junior, Luiz A. L. Rodrigues, Filipe Dwan Pereira, Luciano de Souza Cabral, Newarney T. Costa, Geber L. Ramalho, and Dragan Gasevic. 2025. Automatic Short Answer Grading in the LLM Era: Does GPT-4 with Prompt Engineering beat Traditional Models?. In *Proceedings of the International Learning Analytics and Knowledge Conference*. ACM, 93–103. doi:10.1145/3706468.3706481

[30] Microsoft Corporation. 2012. TypeScript. https://www.typescriptlang.org/

[31] Francesco Maria Molfese, Luca Moroni, Luca Gioffrè, Alessandro Scirè, Simone Conia, and Roberto Navigli. 2025. Right Answer, Wrong Score: Uncovering the Inconsistencies of LLM Evaluation in Multiple-Choice Question Answering. In *Findings of the Association for Computational Linguistics*. Association for Computational Linguistics, 18477–18494.

[32] Phyo Yi Win Myint, Siaw Ling Lo, and Yuhao Zhang. 2024. Harnessing the power of AI-instructor collaborative grading approach: Topic-based effective grading for semi open-ended multipart questions. *Comput. Educ. Artif. Intell.* 7 (2024), 100339. doi:10.1016/J.CAEAI.2024.100339

[33] Sriraam Natarajan, Saurabh Mathur, Sahil Sidheekh, Wolfgang Stammer, and Kristian Kersting. 2025. Human-in-the-loop or AI-in-the-loop? Automate or Collaborate?. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*. AAAI Press, 28594–28600. doi:10.1609/AAAI.V39I27.35083

[34] Bernadette Quah, Lei Zheng, Timothy Jie Han Sng, Chee Weng Yong, and Intekhab Islam. 2024. Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. *BMC Medical Education* 24, 1 (2024), 962.

[35] Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2025. Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the International Learning Analytics and Knowledge Conference*. ACM, 462–472. doi:10.1145/3706468.3706527

[36] Zhengyan Shi, Giuseppe Castellucci, Simone Filice, Saar Kuzi, Elad Kravi, Eugene Agichtein, Oleg Rokhlenko, and Shervin Malmasi. 2025. Ambiguity Detection and Uncertainty Calibration for Question Answering with Large Language Models. In *Proceedings of the Workshop on Trustworthy NLP*. Association for Computational Linguistics, 41–55. doi:10.18653/v1/2025.trustnlp-main.4

[37] Yishen Song, Qianta Zhu, Huaibo Wang, and Qinhua Zheng. 2024. Automated Essay Scoring and Revising Based on Open-Source Large Language Models. *IEEE Trans. Learn. Technol.* 17 (2024), 1920–1930. doi:10.1109/TLT.2024.3396873

[38] Kathrin F Stanger-Hall. 2012. Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. *CBE—Life Sciences Education* 11, 3 (2012), 294–306.

[39] Yuan Tian, Weiwei Cui, Dazhen Deng, Xinjing Yi, Yurun Yang, Haidong Zhang, and Yingcai Wu. 2025. ChartGPT: Leveraging LLMs to Generate Charts From Abstract Natural Language. *IEEE Trans. Vis. Comput. Graph.* 31, 3 (2025), 1731–1745. doi:10.1109/TVCG.2024.3368621

[40] Vue. 2025. Vue.js. https://vuejs.org/ Accessed: June 2025.

[41] Dakuo Wang, Elizabeth F. Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–6. doi:10.1145/3334480.3381069

[42] Ruiqi Wang, Jiyu Guo, Cuiyun Gao, Guodong Fan, Chun Yong Chong, and Xin Xia. 2025. Can llms replace human evaluators? an empirical study of llm-as-a-judge in software engineering. *Proceedings of the ACM on Software Engineering* 2, ISSTA (2025), 1955–1977.

[43] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[44] Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2025. Evaluating Mathematical Reasoning Beyond Accuracy. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*. AAAI Press, 27723–27730. doi:10.1609/AAAI.V39I26.34987

[45] Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2025. Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs. In *Proceedings of the International Learning Analytics and Knowledge Conference*. ACM, 293–305. doi:10.1145/3706468.3706507

[46] ZhenTing Yan, Rui Zhang, and Fei Jia. 2024. Exploring the Potential of Large Language Models as a Grading Tool for Conceptual Short-Answer Questions in Introductory Physics. In *Proceedings of the 2024 International Conference on Distance Education and Learning*. Association for Computing Machinery, 308–314. doi:10.1145/3675812.3675837

[47] Yuheng Zhao, Junjie Wang, Linbin Xiang, Xiaowen Zhang, Zifei Guo, Cagatay Turkay, Yu Zhang, and Siming Chen. 2024. LightVA: Lightweight Visual Analytics with LLM Agent-Based Task Planning and Execution. *IEEE Trans. Vis. Comput. Graph.* (2024), 1–13. doi:10.1109/TVCG.2024.3496112

[48] Yuheng Zhao, Yixing Zhang, Yu Zhang, Xinyi Zhao, Junjie Wang, Zekai Shao, Cagatay Turkay, and Siming Chen. 2024. LEVA: Using large language models to enhance visual analytics. *IEEE Trans. Vis. Comput. Graph.* (2024), 1830–1847. doi:10.1109/TVCG.2024.3368060

[49] Chengbo Zheng, Yuheng Wu, Chuhan Shi, Shuai Ma, Jiehui Luo, and Xiaojuan Ma. 2023. Competent but Rigid: Identifying the Gap in Empowering AI to Participate Equally in Group Decision-Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 351:1–351:19. doi:10.1145/3544548.3581131

[50] Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. ELRA and ICCL, 9340–9351.