Open-ended Structured Question Assessment with Human-LLM Collaboration

Fengyan Lin
School of Software Engineering,
South China University of Technology
Guangzhou, Guangdong, China
daniellin040608@gmail.com

Yanna Lin
Cheriton School of Computer Science,
University of Waterloo
Waterloo, Ontario, Canada
yanna.lin@uwaterloo.ca

Kai Cao
School of Software Engineering,
South China University of Technology
Guangzhou, Guangdong, China
202421045381@mail.scut.edu.cn

Zikun Deng*
School of Software Engineering,
South China University of Technology
Guangzhou, Guangdong, China
Key Laboratory of Big Data and
Intelligent Robot (SCUT), Ministry of
Education
Guangzhou, Guangdong, China
zkdeng@scut.edu.cn

Yi Cai
School of Software Engineering,
South China University of Technology
Guangzhou, Guangdong, China
ycai@scut.edu.cn

Abstract

Open-ended Structured Questions (OSQs) assess not only students' knowledge but also their reasoning and expression. However, grading OSQ requires fine-grained, scoring point–level analysis, which is labor-intensive and difficult to scale. Although recent LLM-based and human–AI collaborative grading systems improve efficiency, they mainly operate at the whole-response level and lack support for point-level inspection, correction, and feedback integration. We present *VeriGrader*, a novel human–AI collaborative system for OSQ grading. It combines chain-of-thought prompting with scoring point– and response-level in-context learning to enable interpretable LLM grading and iterative refinement from instructor feedback. A coordinated multi-view interface supports efficient verification of response segments, matched scoring points, and rationales. We evaluate *VeriGrader* using real course data and a user study with 12 participants. Results show that *VeriGrader* improves both grading efficiency, accuracy, and consistency over the baselines, demonstrating the effectiveness of *VeriGrader* and promoting human–AI collaboration in educational assessment.

CCS Concepts

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

Keywords

LLMs, Education

*Zikun Deng is the corresponding author.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791034

ACM Reference Format:
Fengyan Lin, Yanna Lin, Kai Cao, Zikun Deng, and Yi Cai. 2026. Open-ended Structured Question Assessment with Human-LLM Collaboration. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3772318.3791034

1 Introduction

Open-ended **S**tructured **Q**uestions (OSQs) represent 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, 29]. 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. To grade OSQs, instructors must carefully examine each response and evaluate it against the predefined scoring points to ensure accurate and comprehensive assessment.

However, performing such fine-grained, scoring point–level grading manually is highly challenging due to the diverse, ambiguous, and freely structured responses of the students. This difficulty is further amplified by the need to maintain consistent judgments across responses to ensure fairness. Students express the same idea in highly diverse ways due to differences in language habits and writing styles. Even for the same scoring point, responses can vary greatly in wording and structure. This makes it time-consuming to extract text pieces that may correspond to specific scoring points.

Such diversity also introduces ambiguity when instructors judge whether a specific piece correctly matches the expected semantics of the scoring point. In Fig. 1, responses like "My mom asked me to" or "I'm not embarrassed when I talk to people up close" can be ambiguous to align with the expected scoring points. In large-scale grading, instructors must further maintain consistent semantic judgments across responses and over time to ensure fairness.

To alleviate instructors' workload, research in educational assessment has explored both fully automatic grading algorithms and human–AI cooperative systems. Recent advances in large language models (LLMs) have further motivated the development of automated graders, given their impressive capabilities in natural language processing (NLP) [17, 33, 35, 56]. However, 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 [8, 51, 54]. Furthermore, LLMs continue to struggle with ambiguity [39] and context-dependent complexity in student responses [10], often resulting in inaccurate judgments. To overcome these limitations, recent work has explored human–AI collaboration to combine human agency with AI efficiency in grading tasks, leading to systems for short-answer questions [8, 51] and report assignments [5]. However, these systems primarily focus on the entire response or report and therefore do not extend to the demands of OSQ grading, which requires fine-grained, scoring point–level assessment. OSQ grading asks instructors (or models) to locate, interpret, and evaluate small pieces of a student's response in relation to predefined scoring points—a level of semantic granularity far beyond what current automated algorithms or existing human–AI collaborative systems support. This poses not only challenges for algorithms that aim to automate such fine-grained and point-level evaluation, but also challenges for interaction and workflow design. This includes questions such as how to present fine-grained grading results (e.g., scattered response pieces, matched scoring points, and explanations) for accurate and flexible instructor inspection and correction, and how to let instructor feedback dynamically guide the AI across large-scale grading tasks. These limitations call for a new class of human–AI collaborative systems tailored to OSQs, which pair point-level analysis with interaction designs that support fine-grained inspection, lightweight correction, and reliable large-scale grading.

In this study, we analyzed 141 graded exam papers and interviewed five stakeholders to understand existing OSQ grading workflows, along with the corresponding challenges and needs. The findings highlight users' desire for automated, point-level grading that aligns with their current practices, as well as systems that allow them to inspect and audit AI outputs and provide feedback that can refine the model's grading behavior. Based on the findings, we propose *VeriGrader*, an interactive prototype that unifies LLM grading, instructor verification, and feedback into an instructor–LLM collaborative paradigm for OSQ grading.

First, we optimize the LLM for automated, point-level grading through a prompting-based approach. Particularly, we develop a chain-of-thought prompting strategy that enables the LLM to perform fine-grained, scoring point–level, interpretable, and human-like grading at scale. We further combine this prompting with scoring point– and response-level in-context learning strategies, allowing the LLM to incorporate instructor feedback and iteratively improve its grading across the batch. Second, built upon these algorithmic capabilities and design considerations, we design *VeriGrader* (Figure 4) to support instructors in efficiently inspecting and correcting LLM grading results. After users upload an OSQ, its reference scoring points, and student responses, *VeriGrader* segments each response into pieces, maps these pieces to scoring points with labels of correct, wrong, and unclear, and generates corresponding grading reasons. Users then obtain an overview of LLM's grading for the entire batch and can drill down to verify individual responses. For each response, segmented response pieces, matched scoring points, and the grading reasons are visually presented in coordinated panels with linked highlighting for verification, triggered by user's hovering. Users can then modify the segments, reassign scoring points, and confirm the grading decision. Confirmed scoring point– and response-level grading results are fed back to the LLM as exemplars, guiding re-grading of remaining unverified responses.

A user study with twelve participants, using responses from fifteen students per question collected in real classroom settings, demonstrates that *VeriGrader* improves grading efficiency, accuracy, and consistency over both manual and LLM-only grading.

In sum, our contributions can be summarized as follows:

- We conduct stakeholder interviews and analyze graded exam papers to distill the grading practices and then compile requirements of human–AI collaboration for OSQ assessment.
- We develop a prototype system, *VeriGrader*, the first system supporting instructor-LLM collaborative OSQ grading. *VeriGrader* frees instructors from tedious manual response segmentation, scoring point mapping, and semantic judgment, while retaining their agency to inspect, correct grading results, and guide LLMs, thus achieving credible, reliable, consistent, and efficient OSQ grading.
- Empirical validation of *VeriGrader* in real-world educational contexts, demonstrating its advantages in grading efficiency, accuracy, consistency, and instructor agency, alongside an exploration of future potential and directions for transparent AI support in high-stakes assessment and beyond.

2 Related Work

2.1 Traditional Educational Assessment

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

Assessment Categories. 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 [9, 43]. In contrast, subjective assessments are crucial for gauging higher-order learning outcomes such as critical thinking, analytical reasoning, and creative expression [3, 43]. They require students to synthesize information and construct coherent arguments, and demonstrate deep conceptual understanding, which purely objective methods cannot adequately assess.

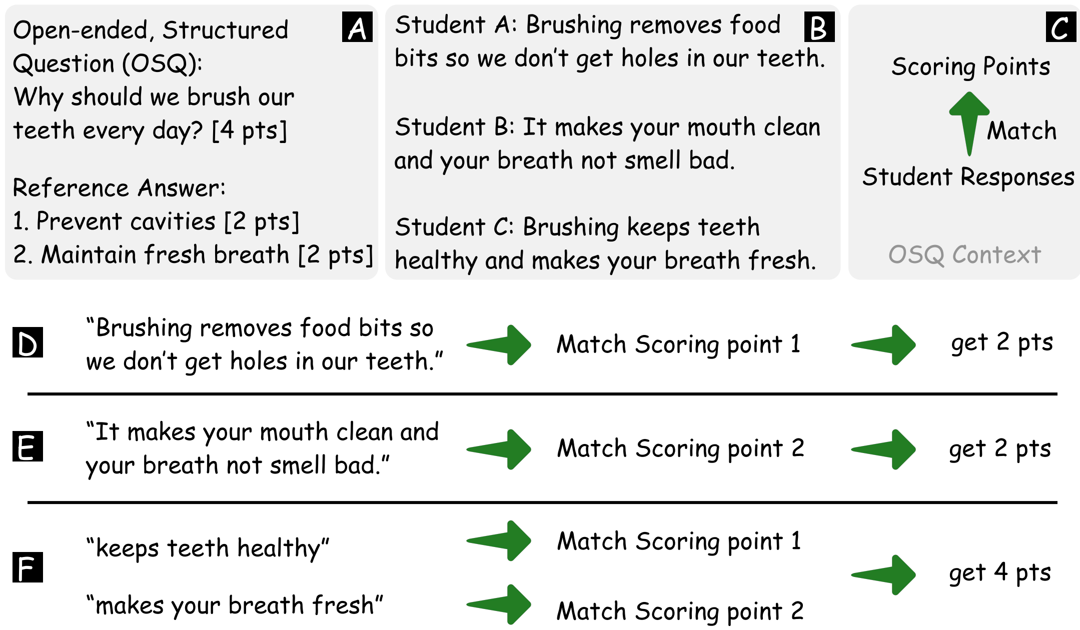

Figure 1: (A) An OSQ and its reference answer with two scoring points. (B) Student responses. (C) Grading process. (D, E, F) Grading results, where the first two responses address only one scoring point each, whereas the third response covers two scoring points.

Grading of subjective assessments. Purely manual grading of subjective assessments presents challenges. First, the workload can be substantial. During high-stakes periods such as final examinations, instructors may be required to grade hundreds of submissions under tight deadlines [13]. Second, while generally reliable, human scoring can still be prone to occasional errors. 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 [19, 28, 37]. 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, 17, 27, 35, 50].

This study focuses on a common type of subjective assessment item, which we term open-ended structured questions (OSQs). OSQs permit open responses, but still require students to organize their textual 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

With the rapid advancement of large language models (LLMs), they are increasingly being integrated into educational assessment workflows, including short-answer evaluation [17, 33], multiple-choice grading [35, 56], essay or report scoring [5, 27, 41, 50], and mathematical problem assessment [2, 49].

Fully automated LLM grading. Many automated grading systems [32, 51, 54] have improved the accuracy of LLM scoring, by using few-shot learning [4], rubric injection [25], or chain-of-thought reasoning [48] to guide the model to generate more reliable scores. In certain settings, such as programming assignments, LLMs have achieved performance on par with human instructors [16, 23, 47]. However, the scores generated by these fully automated systems lack interpretability and transparency. Prior studies have shown that automated models can introduce bias or unfairness in educational assessment [38], for instance, by penalizing creative or culturally diverse responses [6]. A recent survey further reported that 60% of students expressed concerns about AI fairness in grading subjective responses [36], underscoring the importance of maintaining human oversight to guide LLM-based grading.

Human-LLM collaborative grading. Recently, a growing body of work has attempted to introduce reason generation and explanation interfaces [8, 51]. While these approaches improve transparency, the interaction remains largely one-directional: instructors can inspect model outputs but cannot meaningfully inject pedagogical judgment. This design prevents the system from learning from instructor feedback and limits instructor agency in tasks where reliable grading critically depends on contextual understanding and nuanced interpretation. Thus, some systems enabled closer integration and two-way collaborative grading between humans and LLMs. Cohn et al. [8] introduce an active learning–based framework that refines short-answer grading through instructor-provided corrections. Similarly, Chen et al. [5] present a report grading system where instructors review and adjust initial grades and benchmark reports, and the revised inputs are fed back to the AI agent for subsequent regrading.

While these systems are effective, they mainly focus on short-answer questions or report-level tasks. On the AI side, current LLMs cannot perform fine-grained, scoring point-level grading for OSQs, making it difficult to support post-review and audits. On the user side, existing systems do not explicitly account for real-world OSQ grading practices, such as segmenting response pieces, aligning pieces with scoring points, and providing scoring point-level feedback to the LLM. To address these gaps, we propose *VeriGrader*, the first human-AI collaborative OSQ grading system. It enhances LLM grading capabilities and integrates instructor agency through a set of visual and interactive features tailored to OSQ grading practices.

2.3 Human-AI Collaboration

In parallel with rapid AI advances, HCI research has explored how to combine human expertise with AI workflows to enhance reliability, accountability, efficiency, and trust calibration [1, 11, 18, 22, 46]. Based on the default decision authority and the role of AI within the workflow, these approaches can be categorized into three types.

AI-assisted decision making. In this paradigm, AI serves as an advisor, offering recommendations, rationales, or uncertainty estimates, while humans retain full control. In education, this paradigm underlies LLM-based tools that generate adaptive content or suggest code explanations for human review [12, 21, 30].

Human-supervised AI. By contrast, this kind of system proposes a default decision (e.g., a score) that humans may verify or override, prioritizing efficiency under human oversight. Some automated grading systems follow this strategy. For example, Lee and Song [25] developed a system that allows instructors to verify an LLM's grading rationale for short-answer questions. Similarly, Xiao et al. [50] presented grading rationales across overall, content, language, and structural dimensions for essay evaluation.

Human-AI collaboration. In complex tasks, humans and AI tend to assume more equal roles, forming a paradigm of mutual assistance and collaboration [20, 31, 44, 53]. The AI goes beyond providing recommendations [26, 42] to actively participating in decision-making, with human feedback incorporated to iteratively improve its performance. Representative scenarios include visualization design [52], tactical planning[31], and graphic design [55]. Similar collaborative principles have recently begun to emerge in educational assessment, such as Cohn et al.'s [8] and Chen et al.'s [5] approaches discussed before. Instructors not only verify and correct the AI grading results, but their feedback can also be fed back to the AI via active learning or in-context learning.

To the best of our knowledge, we are the first to extend the human-AI collaboration paradigm to the grading of OSQs. This extension is nontrivial, as it addresses challenges that cannot be resolved by directly applying existing paradigms. OSQ grading practices have not been systematically explored, including workflow, interaction and visualization requirements, and instructor needs. Particularly, unlike short-answer or report grading, scoring-point-based OSQ grading involves more than providing comments or determining whether answers are correct. OSQ grading requires first identifying relevant but scattered response pieces from the entire response, accurately mapping them to scoring points with close semantics, and judging the correctness considering diverse expressions of students. Not only should the LLM be optimized for such grading, but the user interface should also support the instructor-LLM communication accordingly.

3 Background Knowledge

Open-ended structured questions (OSQs) are assessment tasks that lie between fully open-ended questions (e.g., essays) and fully structured questions (e.g., multiple-choice). At the task level, each OSQ has clearly defined **scoring points**—the knowledge units or solution steps expected in a complete answer. OSQs can include STEM short-answer questions (e.g., physics explanations, coding tasks, math problems) or reasoning tasks such as word problems, as long as the expected solution elements can be explicitly specified. At the response level, students answer in textual form freely based on their understanding and linguistic habits and are expected to cover as many scoring points as possible. Yet, diverse expressions may pose challenges for consistent, accurate grading. Usually, instructors assess responses against the scoring points, which serve as discrete, semantically complete units that can be individually scored. Students are expected to cover as many scoring points as possible. In this way, scoring points provide a bridge from qualitative responses to quantitative scores, enabling consistent and fine-grained assessment.

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 compiled design requirements that guide the visual and interaction design of the interface.

4.1 Preliminary Interviews

Interviewees. Interviewees included (1) three high school instructors (T1-3), who, on average, had more than eight years of experience in grading and were 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 interviews were 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 interviews were 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 interpretability. 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 OSQs. 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 was 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 was 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 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 agreed 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 imagined 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 aimed to derive insights from past graded exam papers, particularly regarding how instructors interpret student responses and determine scores. Thus, we analyzed 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 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 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 OSQs 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.

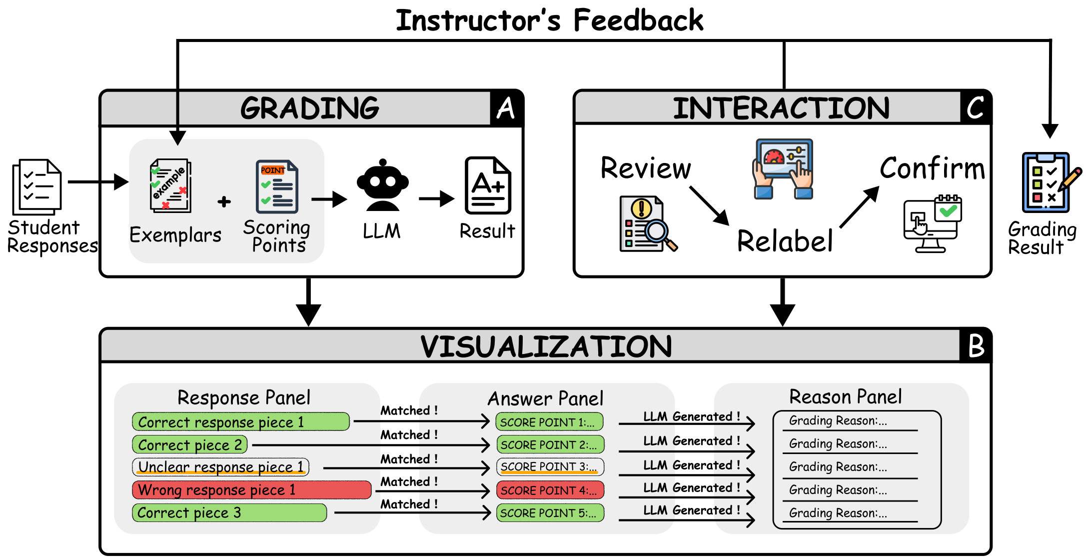

Figure 2: *VeriGrader* workflow. (A) After upload, the system assembles a grading prompt from the question, reference answer, and exemplars, and calls the LLM to grade student responses. (B) The grading results of the LLM are visually exposed to the instructor in three panels. (C) Instructors review 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.

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 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 each student response into pieces, maps these pieces to the scoring points, and categorizes them into three predefined categories (Figure 2-A). These segmentations, categorizations, and grading reasons are mainly visualized in three panels, allowing instructors to review and explain whether student responses are correctly graded (Figure 2-B). During the review process, instructors can refine the segmentations, categorizations, 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

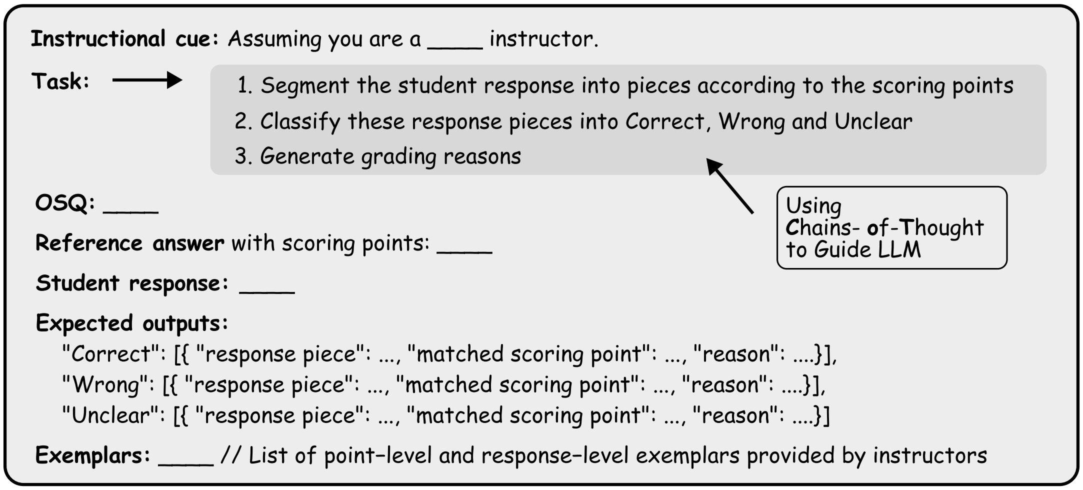

Figure 3: Prompt design for automated grading, including instructional cue, OSQ, reference answers, student response, and exemplars. The LLM is guided to segment responses, classify pieces into Correct/Wrong/Unclear, and generate grading reasons using chain-of-thought reasoning.

generating reasons. From the output of grading each response, we can obtain multiple quadruples <**response piece**, **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 the 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, ranging from 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 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, the navigation view (Figure 4-A) and the grading view (Figure 4-B).

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

In the navigation bar (Figure 4-A1), instructors can switch questions via tabs and navigate responses through numbered buttons. Each button encodes grading status: active response (blue), LLM-graded response (green), and instructor-verified exemplar or grading standards (yellow). Confirmed results are outlined in blue. To help instructors prioritize responses that may contain unreliable LLM grading results, *VeriGrader* assigns each response a quality priority based on two shallow indicators: UR Rate measures the proportion of entities classified as unclear, reflecting uncertainty in LLM classification or low response clarity, while Overlap Rate examines whether response pieces overlap within a student response, signaling inconsistent boundary detection. Responses are labeled high, medium, or low priority:

- High Priority (red dot): Overlap Rate > 0, or UR Rate > 50%.
- Medium Priority (orange dot): UR Rate between 25–50%.
- Low Priority (no dot): all indicators fall within the normal range.

Figure 5 provides a visual summary of these indicators. Response can be sorted accordingly, with exemplars always placed first. This sorting directs attention to responses most likely to require verification. Summary statistics for each question (highest, lowest, average score) are shown in Figure 4-A2.

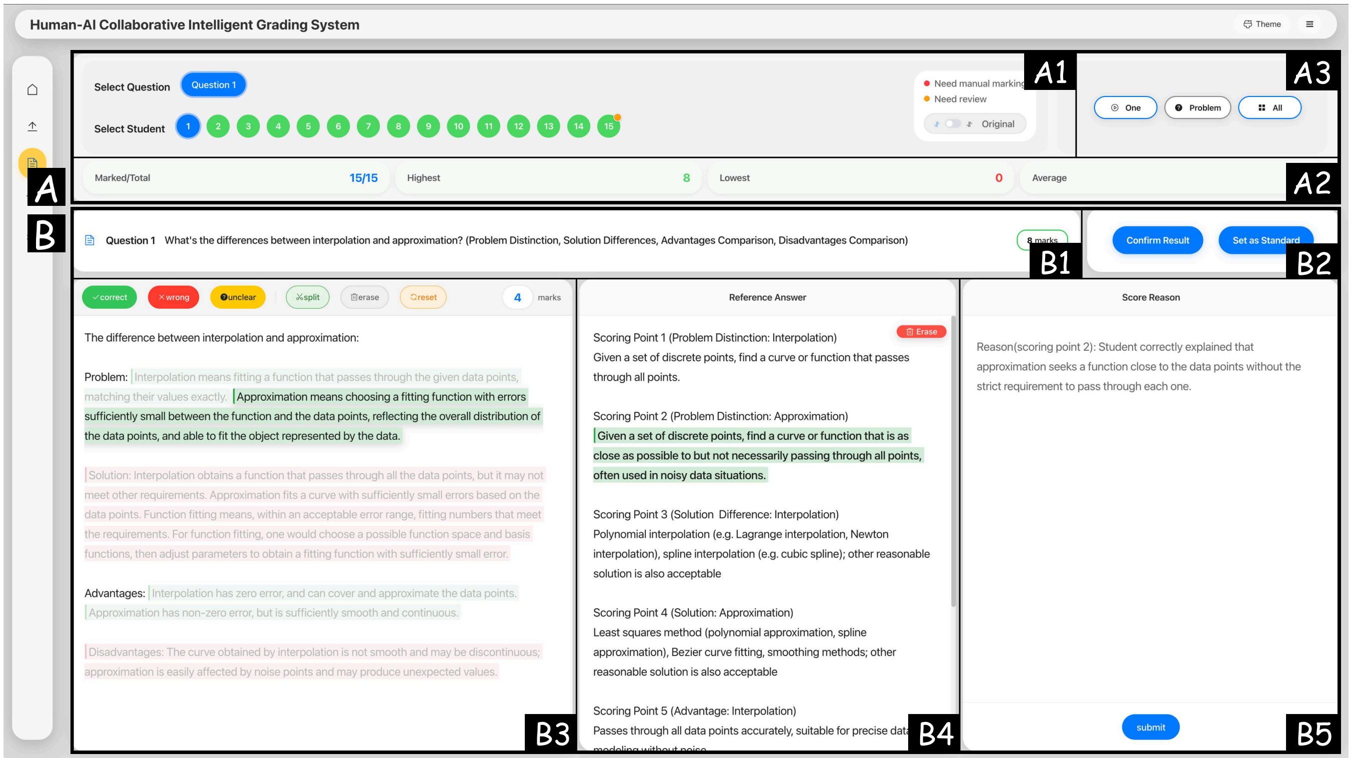

Figure 4: *VeriGrader* System Interface. The interface comprises two views: a navigation view (A) and a grading view (B). In the navigation view, A1 provides a tab-based question selector and numbered response buttons, with a toggle for quality-based sorting; A2 summarizes performance with min/avg/max 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 point aligned to the selected response piece, while the reason panel (B5) displays an editable, LLM-generated grading reason for the current selection.

Figure 5: Visual encoding and priority indicators

The navigation view also provides three LLM regrading options (Figure 4-A3): regrade only the current response ("One" button), regrade medium/high-priority responses ("Problem" button), or regrade all unconfirmed responses ("All" button). These operations help instructors quickly apply newly added exemplars to improve grading quality.

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.

At the top, the selected question and its total score are shown (Figure 4-B1), helping instructors recall the question. Two buttons are provided (Figure 4-B2). "Confirm" finalizes the current grading result, whether generated by the LLM or refined manually. "Set as Standard" marks the current result as an exemplar. The corresponding response button in the navigation bar turns yellow, the response is locked from further edits, and it is added to the exemplar set as a standard that guides subsequent LLM grading, ensuring authoritative scoring.

The lower section contains three linked panels:

- **Response Panel** displays the student response decomposed into pieces detected by the LLM (Figure 4-B3). Pieces are color-coded by category: green for correct, red for wrong, and yellow for unclear. Overlapping spans are visualized with consistent ordering. The score updates automatically when pieces are modified.
- **Answer Panel** shows the reference answer with highlighting and linking support (Figure 4-B4), enabling instructors to quickly relate response pieces to scoring points.
- **Reason Panel** presents the model-generated reason for the selected piece (Figure 4-B5), which instructors may view, edit or regenerate.

Hovering over a response piece highlights the corresponding scoring point and displays its rationale, enabling instructors to inspect and correct LLM outputs. Instructors may reassign categories, remove irrelevant pieces, remap pieces to different scoring points, or add new pieces that the LLM missed. All updates propagate across panels to maintain consistency.

VeriGrader gradually improves by incorporating instructor feedback through exemplars. Instructors can submit individual point-level exemplars directly from the reason panel. Besides, a fully graded response can be marked as a response-level exemplar that aggregates all its pieces. All these exemplars are then used to guide subsequent LLM regrading. Over time, this process allows *VeriGrader* to adapt to the instructor's grading style, producing automated scores that are increasingly consistent and aligned with instructional intent.

More details can be found in the usage scenarios below and in the accompanying video.

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 [45] and TypeScript [34], 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 was 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* to perform the grading process.

Daniel uploaded the responses and reference answers to *VeriGrader*. Within 30 seconds, the LLM completed 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 was simultaneously highlighted. Based on his expertise, Daniel determined that both *unclear* pieces were actually wrong, as "*multiple forms of interpolation*" and "*multiple forms of Approximation*" were too vague to correspond to any valid scoring points. He then relabeled them as wrong (Figure 6-A2). The system automatically regenerated the associated grading reason. Afterwards, he reviewed 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 cannot 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 response-level 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 invoked 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 response-level exemplars within the system.

The third response presented a new case (#*id* = 2). 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 may cause oscillations" as the scoring point (Figure 8-C). The system initially did 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 had 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 refined the entire grading process. After several iterations and accompanied by reviewing, Daniel completed grading all responses. Each response was associated with clear and explicit scoring points and reasons. *VeriGrader* not only improved efficiency but also internalized Daniel's grading standards and pedagogical principles, thereby ensuring consistency and fairness in the final outcomes.

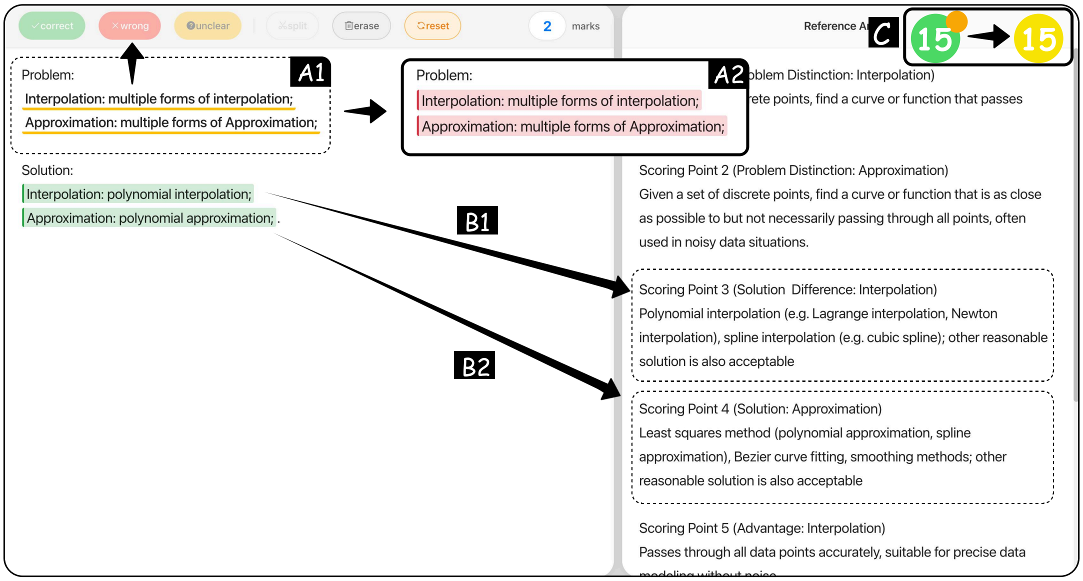

Figure 6: Resolving unclear response pieces. Two initially unclear response pieces (A1) were manually reassigned to the *wrong* category (A2). The system retained the two remaining correct response pieces (B1–B2) with their associated reference answers. Finally, the instructor confirmed the response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

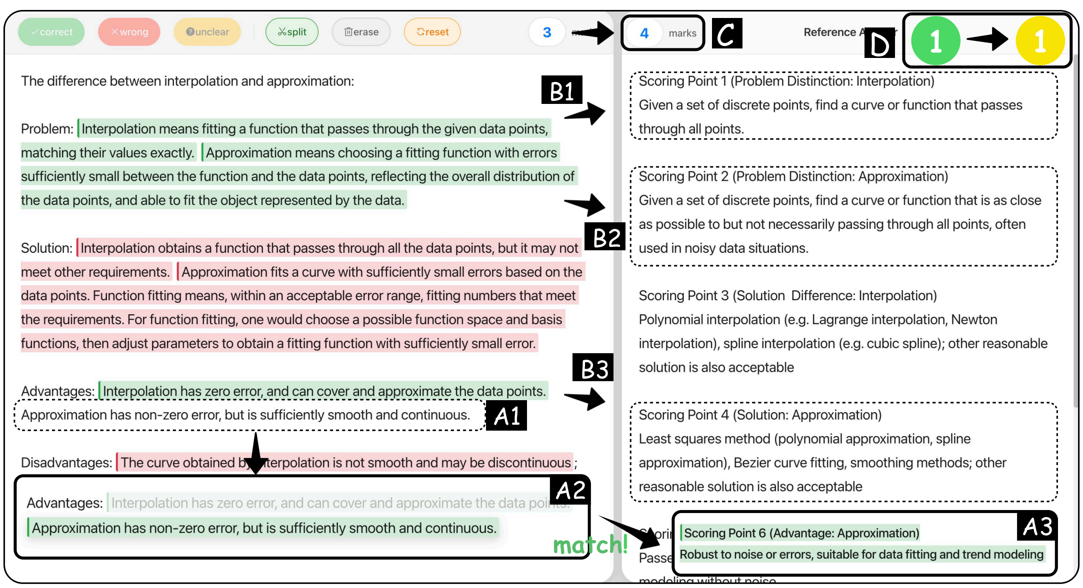

Figure 7: Add the missing scoring points. An initially unscored response piece (A1) was manually marked correct and highlighted as A2, prompting the system to match the corresponding scoring point from the reference answer (A3) and update the total score (C). Together with the three previously recognized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

7 User Study

To evaluate the effectiveness of *VeriGrader*, we designed a controlled study. In this study, we compared *VeriGrader* grading of open-ended structured questions (OSQs) with a manual grading baseline system, which was developed to simulate the traditional manual grading process and ensure a fair comparison. Based on this setup, we formulated four research questions:

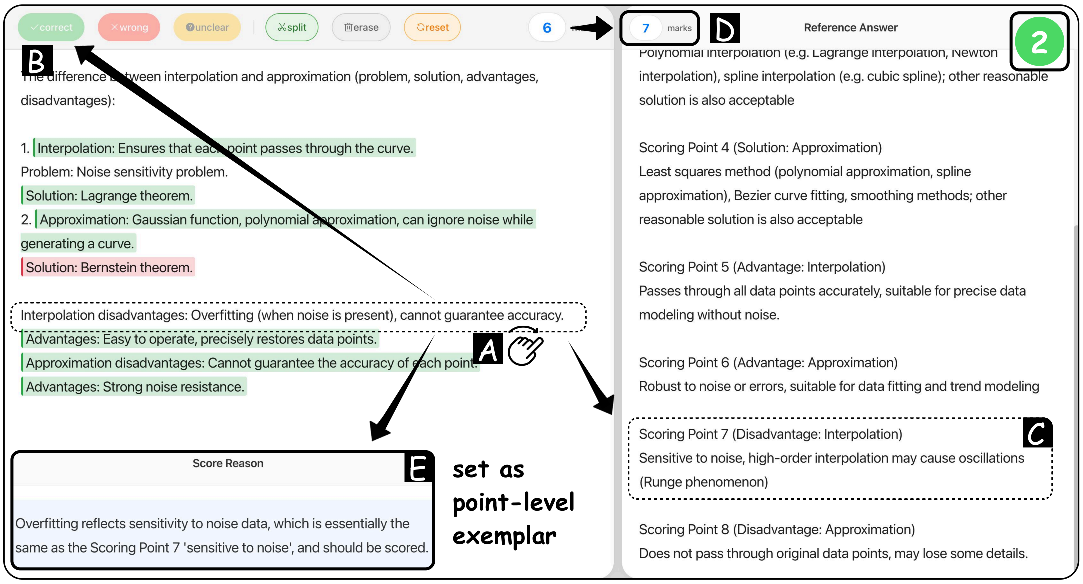

Figure 8: The instructor highlighted a student's response piece by dragging the mouse (A). He then assigned the appropriate category (B) , triggering the system to automatically invoke the model. As a result, a corresponding reference answer (C) and grading reason were generated, and the score was updated (D). To ensure consistency in future grading, the instructor further edited the rationale and saved it as a point-level exemplar (E).

- RQ1: Can *VeriGrader* outperform manual grading in terms of accuracy and efficiency?
- RQ2: Under absolutely no human intervention, can LLMs achieve high grading performance?
- RQ3: Does introducing few-shot exemplars lead to improvements in the model's grading performance?
- RQ4: How does *VeriGrader* perform consistently across multiple instructors?

7.1 Study Setup

Dataset. We compiled a small-scale yet representative dataset that integrates carefully designed OSQs, instructor-authored reference answers with scoring points, and real-world student responses. The dataset contains two OSQs, selected for their coverage of distinct knowledge domains, namely **numerical methods** and **data structure**, 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 captured 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 12 participants (U1–U12; 9 males and 3 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, Gemini, DeepSeek, and Segment Anything.

Although the sample size is modest, recruiting participants with prior grading experience was challenging 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 *VeriGrader*, removed its LLM-assisted functions, and replaced it with manual annotations, resulting in a manual grading system. Specifically, in the manual grading system, participants 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* with the manual grading system (**RQ1**), each participant was required to grade responses under Setting **S1** and under Setting **S2**, respectively. To examine the effectiveness of incorporating user feedback (**RQ3**), 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 structure 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

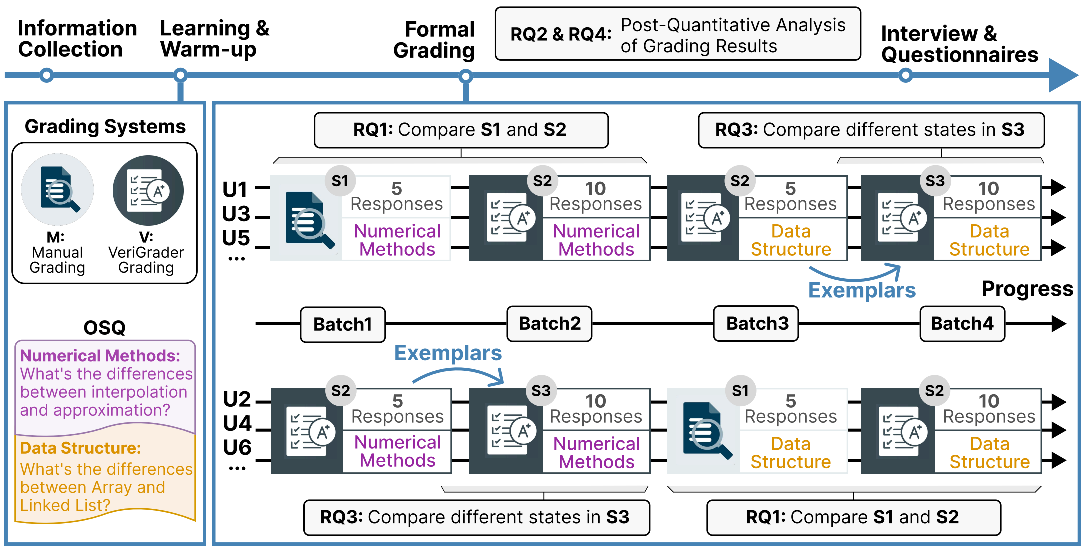

Figure 9: The experiment procedure consisted of four sequential steps: (1) Information Collection, obtaining consent and gathering participant background; (2) Learning and Warm-up, introducing system functions and practicing with sample OSQ; (3) Formal Grading, performing detailed grading of student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

experimental procedure. After the experiments, we performed post-quantitative analysis of the grading results to address **RQ2** and **RQ4**.

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 provided 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 was 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 four research questions simultaneously and reduce the learning effect, we designed the experimental pipeline shown in Figure 9. Each participant goes through both the manual grading system (Setting **S1**) and *VeriGrader* (Setting **S2**). The order of the systems varies for different participants. Particularly, the grading results generated by participants using *VeriGrader* in Setting **S2** will be immediately used for the process in which the participant 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 adopted a multi-expert validation procedure. We invited three domain experts with extensive experience to perform fine-grained manual grading on all student responses. Following real-world grading practices, the experts carefully compared each student response r_i against the reference answers and identified which scoring points it covered. When disagreements occurred, the experts engaged in focused discussion sessions in which they jointly reviewed the contested responses, explained their interpretations, and resolved discrepancies through consensus. Ultimately, each response r_i was assigned a unified ground truth represented as a set of tuples <response piece, scoring point, category>.

Accuracy. Objective questions (true/false and multiple-choice) are typically evaluated by the fraction of correctly answered items. Subjective questions (short-answer or open-ended) have been assessed using various metrics in prior work (e.g., AUC and RMSE).

In contrast, our approach produces fine-grained, scoring point-level scores, thus we can evaluate each scoring point against the ground truth, providing not only accurate but also more interpretable, human-aligned assessment. For each student response r_i, each grading result by a participant alone, the LLM alone, or a participant using *VeriGrader* is represented as Result_i = <response piece, scoring point, category>. We compared the grading results against the ground truth. Let n denote the total number of student responses. *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 participant first views the response to the moment they move on to the next one. We used the *average time spent per OSQ* (minutes/OSQ) for a batch of responses as the efficiency measure.

Piece Matching Rate (PMR). We also want to measure the proportion of LLM-segmented pieces that match ground-truth pieces under a word-level Jaccard similarity threshold J.

$$\text{PMR} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\Big(\max_{g \in G_i} J(r_i, g) > J\Big),$$

where N is the number of ground-truth pieces, r_i is the i-th LLM piece, G_i is the set of ground-truth pieces for the same response, $J(r_i, g)$ is the word-level Jaccard similarity between r_i and g, J is the matching threshold, and $\mathbf{1}(\cdot)$ is the indicator function.

PMR characterizes the alignment between LLM-generated segmentation and human segmentation. This alignment defines the unit of analysis for grading, which directly affects both LLM grading and instructors' efficiency in reviewing and correcting scores. When segments are misaligned (low PMR), the model may assign scores to incorrect pieces, and instructors must add, split, or merge segments to correct these errors, increasing workload. Conversely, higher PMR reduces segmentation mismatches, improving the reliability of model grading and minimizing the need for manual corrections. Consequently, PMR primarily shapes instructors' correction effort and interaction efficiency, while also indirectly influencing downstream grading performance within the hierarchical workflow.

Consistency. To assess the consistency among instructors (participants) in scoring student responses, we employ multiple complementary reliability measures.

- *Weighted Fleiss' Kappa (κ_w).* At the total score level, we employ Weighted Fleiss' Kappa [14]. Unlike the standard Fleiss' Kappa, which treats all disagreements equally, the weighted variant incorporates a weight matrix w_{ij} that assigns smaller penalties to minor ordinal differences and larger penalties to substantial deviations. In our study, we adopt quadratic weights for the difference between scores. Together, these choices yield a single reliability statistic that captures how consistently the 12 participants scored the full set of student responses.
- *Intraclass Correlation Coefficient (ICC).* To further evaluate the consistency among instructors in assigning total scores to each response, we also employ the Intraclass Correlation Coefficient (ICC) [40]. It provides a complementary statistic, supporting and reinforcing the insights obtained from Weighted Fleiss' Kappa.
- *Gwet's Agreement Coefficient 1 (AC1).* At the scoring-point level, we employ Gwet's Agreement Coefficient 1 (AC1) [15]. Unlike traditional Kappa statistics, AC1 is less sensitive to prevalence and marginal imbalance, providing a more stable measure of consistency when most instructors give the same rating. In our study, it is used to quantify whether multiple participants consistently mark the presence or absence of each scoring point across all responses. This metric yields a single reliability statistic that reflects the inter-rater consistency for individual scoring points, complementing the overall score consistency measured by Weighted Fleiss' Kappa.

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** in consecutive batches of responses (Batches 1–2 and Batches 3–4). Setting **S2** reflects the common human–AI collaborative workflow in *VeriGrader*, while Setting **S1** serves as the manual grading baseline.

The results, detailed in Table 1 and shown in Figure 10, indicated that *VeriGrader* outperformed the manual system in both efficiency and accuracy, and these advantages can be achieved simultaneously. On average, the grading time per response decreased by 56.3% (from 3.82 ± 1.4 to 1.67 ± 0.5 minutes), accompanied by gains in accuracy (86.26 ± 10.2% to 91.86 ± 5.0%).

We analyzed Efficiency and Accuracy using linear mixed-effects models (LMMs), with Question (OSQ1 and OSQ2), System (Manual and *VeriGrader*), and their interaction as fixed effects and Participant as a random intercept. For Efficiency, the model revealed a significant main effect of System (β = -2.783, p < .001), indicating that *VeriGrader* substantially reduced completion time. The main effect of Question was not significant (p = .179). Importantly, the Question × System interaction was significant (β = 1.267, p = .033), suggesting that the time difference between systems varied across questions. For Accuracy, the model showed a significant main effect of System (β = 8.717, p = .034), indicating that *VeriGrader* improved response accuracy. Neither the main effect of Question (p = .578) nor the interaction (p = .284) reached significance. Together, these results indicate that OSQ1 and OSQ2 exhibit comparable grading

Table 1: Comparison of Setting S1 and Setting S2 on OSQ1 (data structure) and OSQ2 (numerical methods). Across all participants, *VeriGrader* substantially reduced grading time 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.

	OSQ1: Data Structure						OSQ2: Numerical Methods						
	U2	U4	U6	U8	U10	U12	U1	U3	U5	U7	U9	U11	**Avg**
F1_Manual [S1]	90.6	62.2	94.5	84.0	85.7	92.6	71.8	97.9	82.6	89.8	89.8	93.6	**86.26**
F1_VeriGrader [S2]	95.7	94.8	93.2	94.3	88.2	95.7	79.5	86.3	90.3	93.8	94.6	95.9	**91.86**
Δ	+5.1	+32.6	-1.3	+10.3	+2.5	+3.1	+7.7	-11.6	+7.7	+4.0	+4.8	+2.3	**+5.6**
min/OSQ_Manual [S1]	3.4	6.0	4.6	4.8	3.0	3.6	2.0	2.4	2.2	6.4	4.4	3.0	**3.82**
min/OSQ_VeriGrader [S2]	1.0	1.6	1.6	1.7	1.2	1.6	1.6	1.4	1.4	3.0	2.2	1.7	**1.67**
Δ	-2.4	-4.4	-3.0	-2.9	-1.8	-2.0	-0.4	-1.0	-0.8	-3.4	-2.2	-1.3	**-2.15**
%Δ	-70.6	-73.3	-65.2	-64.6	-60.0	-55.6	-20.0	-41.7	-36.4	-53.1	-50.0	-43.3	**-56.3**

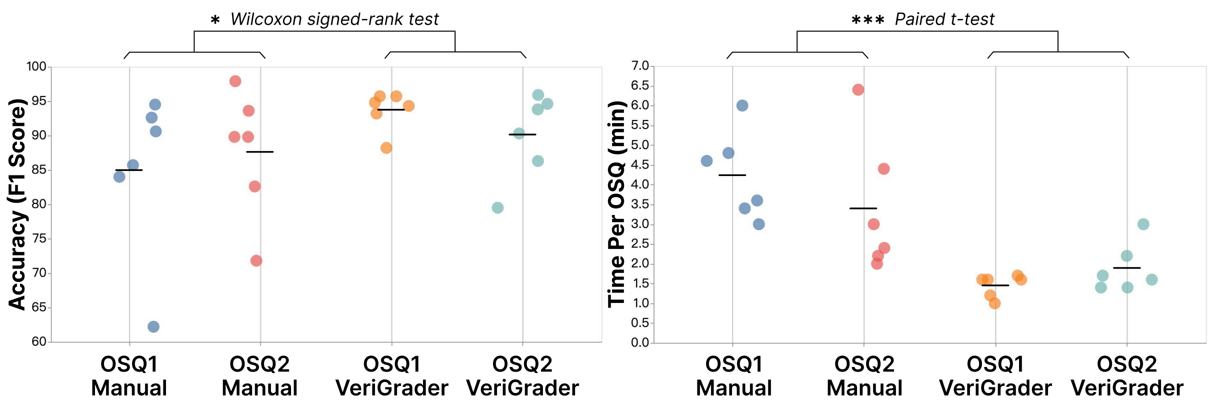

Figure 10: Comparison of grading accuracy (F1 Score) and efficiency (time per OSQ) between manual grading and *VeriGrader* for OSQ1 (data structure) and OSQ2 (numerical methods). In the jitter plots, each dot represents one participant, with the black line indicating the mean. The left panel reports F1 accuracy, and the right panel reports average time per response. Results demonstrate that *VeriGrader* converges participants' accuracy to a consistently high level while simultaneously converging time per OSQ to a substantially lower level. Significance notation: $*$ indicates $p < .05$; $***$ indicates $p < .001$.

difficulty, and any observed differences in performance can be attributed to system effects rather than differences in question difficulty.

We compared the performance of *VeriGrader* and Manual in terms of *Efficiency* and *Accuracy*, treating each participant as a paired observation. These results are based on all 12 participants. For *Efficiency*, the difference between systems was approximately normally distributed (Shapiro-Wilk test: $W = 0.977$, $p = 0.971$), allowing the use of a paired t-test. The paired t-test revealed that *VeriGrader* significantly reduced grading time compared to Manual ($t(11) = 6.30$, $p < 0.001$), with mean completion times of 3.82 minutes and 1.67 minutes for Manual and *VeriGrader*, respectively (mean difference 2.15 minutes). For *Accuracy*, the difference between systems was not normally distributed (Shapiro-Wilk test: $W = 0.813$, $p = 0.013$), so we applied the Wilcoxon signed-rank test. Results indicated that *VeriGrader* yielded significantly higher grading accuracy than Manual ($W = 12.0$, $p = 0.034 < 0.05$), with mean accuracies of 86.26% and 91.86% for Manual and *VeriGrader*, respectively (mean difference 5.60%). These results demonstrate that *VeriGrader* both accelerates grading and improves accuracy relative to Manual.

It is worth noting that, although *VeriGrader* generally improved grading performance, there were notable individual differences among participants. For instance, U3 and U6 experienced a slight decrease in accuracy when using *VeriGrader* compared to Manual. In contrast, U4 exhibited substantial improvements in both grading speed and accuracy with *VeriGrader*. These individual variations suggest that while *VeriGrader* generally improves efficiency and accuracy, user-specific factors such as strategy, familiarity, or interaction style may influence the benefit, highlighting the need for adaptable and personalized AI-assisted grading tools.

RQ2: The LLM itself achieves high accuracy using the prompting technique. We evaluated the LLM performance using selected grading results drawn from four batches, all graded without exemplars or human intervention. For each of OSQ1 and OSQ2, there are 15 responses. Each of the 12 participants triggered one round of LLM-only grading over the same 15 responses, yielding a total of 12×15 LLM-graded results per OSQ. We averaged the accuracy to estimate the LLM's standalone grading performance while mitigating output variability. The LLM achieved an average accuracy of 89.5% on OSQ1 (data structure) and 89.2% 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.

Beyond grading accuracy, we also measured the LLM's segment quality using the piece matching rate (PMR) under different matching thresholds. As summarized in Table 2, the LLM achieved a reasonably high segment correctness across both OSQs. For example, a threshold of $J > 0.7$ means that the LLM-segmented piece and the ground truth segment share at least 70% of their words. Under this criterion, the LLM correctly matches about 70.1% of pieces in OSQ1 and 74.1% in OSQ2, i.e., LLM can capture at least 70% semantic units in student responses. It supports subsequent grading behaviors, for example, verify whether the response piece semantically matches the scoring point, thereby reducing manual effort. We further examined whether response-level exemplars improved segmentation. We evaluated segmentation on batches 2 and 4 after the LLM used grading results from batches 1 and 3 as five-shot exemplars (Table 2). Results showed that with just five exemplars, PMR improved, suggesting that few-shot prompting helped stabilize segmentation. Although the improvement is modest, it reduces the number of segmentation errors that require manual correction, thereby lowering instructors' grading effort. Even small gains therefore lead to noticeable reductions in instructor workload when grading large numbers of responses.

To further examine the sources of error, we analyzed the LLM's incorrect grading results. We found that the LLM sometimes under-segments information-dense sentences containing multiple scoring points, treating them as a single response piece. This is biased against students who are used to writing concisely, which will lead to them receiving lower scores. In these instances, instructors refined the segmentation through interaction, enabling precise, point-level scoring. The LLM may also miss implicit ideas or informal phrasing, raising concerns about potential bias against students whose responses deviate from standard academic phrasing. In addition, although the unclear category prompts the LLM to defer genuinely ambiguous response pieces to the instructor, the LLM still occasionally forces borderline cases into correct or wrong, reflecting overconfident classification. These patterns suggest opportunities for improving future LLM-based grading.

RQ3: 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 improved accuracy in most cases, with effects more pronounced in OSQ1 (data structure). Moreover, when combined with final human review, accuracy gains were further reinforced, suggesting that human–AI collaboration can enhance grading reliability beyond model-only improvements.

Across all participants, average accuracy improved by 2.7% after introducing few-shot exemplars and approximately 5.6% after final review relative to the initial baseline, indicating consistent gains throughout the collaborative workflow. These results demonstrate that few-shot learning generally strengthens the model's grading ability, while human review consistently elevates accuracy to a high and reliable level.

Most participants showed steady, stage-by-stage accuracy gains, while high-baseline participants (e.g., U11 and U12) occasionally exhibited small fluctuations that were corrected during the final review. For example, U9, beginning at 91.5%, improved to 94.1% with few-shot learning but exhibited a slight fluctuation to 93.6% in the final review stage—likely reflecting slight differences in judgment between the participant and the ground truth. This pattern suggests that when baseline performance is already high, few-shot exemplars may introduce minor perturbations, but human review can stabilize and elevate accuracy toward optimal levels.

RQ4: *VeriGrader* fosters consistent grading practices across instructors. We collected grading results from the second and fourth batches to evaluate inter-rater consistency. For each OSQ, 10 responses were graded by all 12 participants, yielding 120 results per question. An equal number of LLM-only graded results from the same two batches were collected for comparison.

Across OSQ1 and OSQ2, the LLM alone achieved substantial ($\kappa_w = 0.645$) and almost perfect ($\kappa_w = 0.912$) consistency, respectively, according to the interpretation guidelines of Landis and Koch [24]. With *VeriGrader*, consistency further improved to $\kappa_w = 0.771$ and $\kappa_w = 0.916$, respectively.

In addition, we report Intraclass Correlation Coefficients, treating instructors as random effects and student responses as fixed effects. Using a two-way random-effects model, ICC(2,1), the LLM alone reached moderate and excellent reliability ($\mathrm{ICC}(2,1) = 0.670$ and 0.920), while *VeriGrader* improved these values to 0.790 and 0.924, indicating good and excellent reliability according to established guidelines [7]. To capture more fine-grained patterns, we further evaluated Gwet's AC1 at the level of individual scoring points (Table 3). Introducing human supervision via *VeriGrader* not only increased consistency on lower-consistency points in OSQ1 but also stabilized variability across scoring points in OSQ2, demonstrating its role in mitigating model errors while preserving strong baseline consistency. Moreover, to provide a human-only reference, we computed all three consistency metrics based on a small set of manually graded responses under Setting **S1** (5 responses graded by 6 participants). For OSQ1 and OSQ2, κ_w was 0.485 and 0.812, the corresponding ICC(2,1) was 0.559 and 0.846, and the average Gwet's AC1 at the scoring-point level was 0.613 (SD = 0.208) and 0.716 (SD = 0.160), respectively. Across all metrics, the consistency achieved by human-only grading remained lower than that obtained with *VeriGrader*, thereby providing a contextual baseline for interpreting the consistency gains enabled by the human–AI collaborative workflow. While the sample size is smaller and the conditions are not strictly controlled, this comparison still serves as a useful reference for contextualizing the observed consistency gains.

Table 2: Jaccard-based piece matching rates (PMR) with LLM alone and with few-shot exemplars introduced.

Description		OSQ1 J>0.6	OSQ1 J>0.7	OSQ1 J>0.8	OSQ1 J>0.9	OSQ2 J>0.6	OSQ2 J>0.7	OSQ2 J>0.8	OSQ2 J>0.9
LLM's PMR		0.744	0.701	0.652	0.578	0.767	0.741	0.706	0.618
LLM's PMR using 5 exemplars	Before	0.738	0.700	0.674	0.616	0.829	0.813	0.779	0.689
	After	0.773	0.735	0.699	0.609	0.840	0.822	0.790	0.737
	Δ	+3.5%	+3.5%	+2.5%	-0.7%	+1.1%	+0.9%	+1.1%	+4.8%

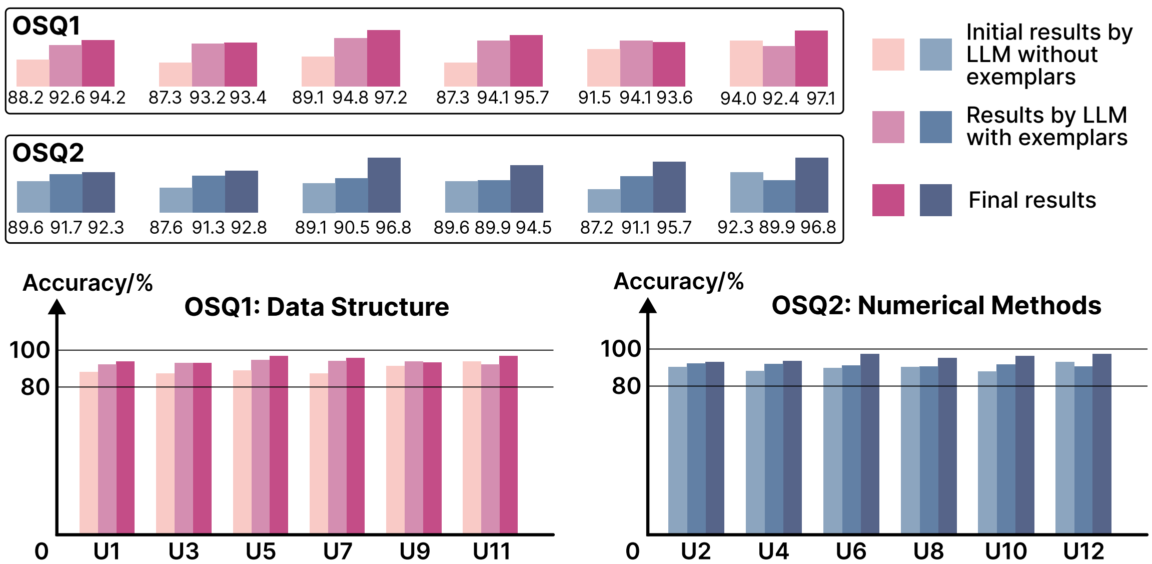

Figure 11: Grading accuracy with and without few-shot exemplars, and with additional human review and refinement, across OSQ1 (data structure) and OSQ2 (numerical methods). The introduction of few-shot exemplars improved accuracy for the vast majority of participants (10 out of 12), with average gains of 3.9% on OSQ1 and 1.5% on OSQ2. Although few-shot learning introduced minor fluctuations in rare high-baseline cases (U11, U12), final human review universally elevated accuracy to an average of 95% across both OSQs. These findings demonstrate that few-shot learning is particularly effective for complex, lower-baseline tasks, while final human review and refinement provide an additional reliability layer.

In sum, these results suggest that the LLM with carefully designed prompts alone can yield high baseline consistency, and the interactive human-AI workflow enabled by *VeriGrader* further reduces variability, enhances consistency, and supports fairer assessments, while maintaining instructors' professional judgment.

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. Core feedback can be summarized as follows:

F1: Showing a "cautious trust" in the LLM capabilities. All participants (12/12) acknowledged the understanding and classification capabilities of LLM in grading (Tailored Q1, $Avg = 6.00$). However, their views on reliability varied (Tailored Q2, $Avg = 5.00$). 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 may 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.*" As U9 explained, "*I trust its first pass, but I still want to verify them.*" 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. 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*

Table 3: Scoring point-level Inter-Rater Consistency (Gwet's AC1) with LLM alone and with *VeriGrader*

	Point 1	Point 2	Point 3	Point 4	Point 5	Point 6	Point 7	Point 8	Avg	SD
OSQ1 (LLM alone)	0.927	0.705	0.755	0.794	0.746	0.822	0.654	0.610	**0.752**	**0.093**
OSQ1 (*VeriGrader*)	0.985	0.911	0.719	0.805	0.830	0.962	0.924	0.877	**0.877**	**0.083**
OSQ2 (LLM alone)	0.977	0.791	0.879	0.620	1.000	0.786	0.684	0.881	**0.827**	**0.125**
OSQ2 (*VeriGrader*)	0.920	0.772	0.749	0.636	1.000	0.610	0.779	0.941	**0.801**	**0.133**

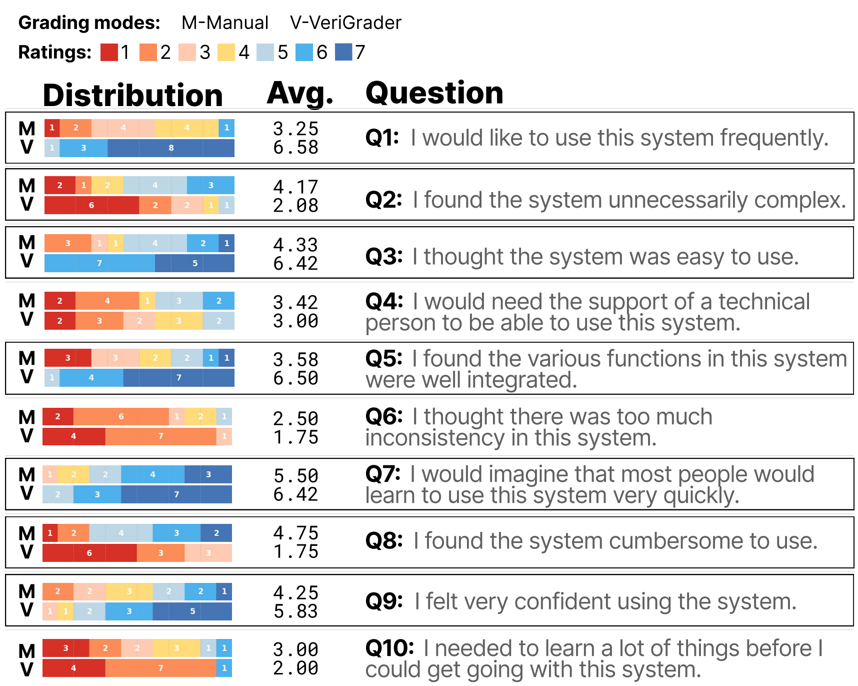

Figure 12: Results of the SUS questionnaire. Boxes indicate cases with significant differences under the Wilcoxon test ($p < .05$).

that correspond to the same scoring point and assign credit only once, while the system can complete this process automatically." The tailored questionnaire results confirmed this advantage: Tailored Q3, *Avg* = 6.00 (efficiency improvement) and Tailored Q4, *Avg* = 6.25 (more accurate capture of scoring points). In this way, instructors could focus more on judgment rather than retrieval.

F3: LLM-generated reasons enhance interpretability and confidence. Participants widely acknowledged the LLM-generated grading reasons. The tailored questionnaire results indicated high agreement that the highlighted pieces and corresponding explanations were reliable in understanding LLM's decisions (Tailored Q1, *Avg* = 6.00). As U10 noted: "*The reason panel helped me know why the model gave this score.*" Others reported that the explanations improved interpretability in ambiguous cases. U7 remarked that when a piece first marked as unclear, "*The model did a great job of explaining why it thought this was ambiguous, which was very helpful for me to make a further judgment.*" Several participants further commented that alignment between their own judgment and the model-generated reason boosted their confidence in the system. U11 stated: "*When I saw the LLM's reasoning was the same as mine, it strongly boosted my confidence.*" However, some participants noted limitations in explanation depth. For example, U4 commented that "*sometimes the explanation felt too generic,*" pointing to opportunities to further improve explanation granularity.

F4: 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 example, U1 said: "*I clearly knew where the score came from.*" The SUS results (Figure 12) also supported this finding. *VeriGrader* achieved much higher ratings than the manual system on usability (Q3, *Avg*=6.42 vs. 4.33) and function integration (Q5, *Avg*=6.50 vs. 3.58). For willingness to use frequently, the manual system received a low score of 3.25, whereas *VeriGrader* reached 6.58.

F5: 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

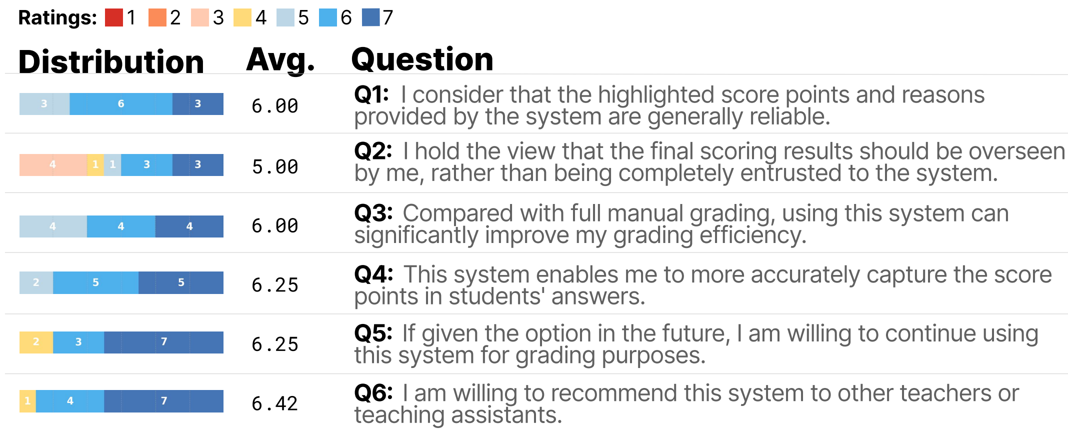

Figure 13: Participants' responses to the 7-point Likert-scale questionnaire items assessing. Higher scores indicate stronger agreement.

using the system (Q5, *Avg* = 6.25) and willingness to recommend it to others (Q6, *Avg* = 6.42). 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, *Avg* = 6.00), and enabled more accurate capture of scoring points (Q4, *Avg* = 6.25). 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 F1.

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.*" U11 also described, "*It felt like co-grading, I provide expertise, the model applies it systematically.*" 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 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. While aligned with general guidelines for human–AI interaction, our principles extend them by addressing domain-specific concerns of auditability, instructor authority, and scoring point–based reasoning that are central to high-stakes grading.

First, a **transparency-first** principle is essential. Beyond explainability, educational assessment requires a stronger form of auditability: instructors must be able to trace each LLM decision to specific scoring points and textual evidence. *VeriGrader* makes this possible through fine-grained response segmentation and scoring point mapping, ensuring that instructors retain full control and can validate each intermediate decision.

Second, systems should **support calibrated trust and avoid instructor overreliance**. General HAI guidelines encourage appropriate reliance, yet the grading context demands a process through which instructors can gradually build trust as the system adapts to their expectations. We observed that *VeriGrader* fosters such calibrated trust: instructors begin cautiously, but as the LLM internalizes their grading preferences through iterative correction, trust and accuracy increase together—addressing risks of over-automation.

Third, **fine-grained bidirectional feedback** enables effective collaboration. Educational assessment introduces a distinctive feedback mechanism. Instructor corrections directly adjust how the LLM interprets and applies scoring points in future cases. In turn, LLM-generated scoring point matches provide structured entry points that streamline human inspection. This tight, fine-grained feedback cycle differentiates assessment workflows from typical human–AI interactions.

When extending AI-assisted systems to new application domains, maintaining **human professional authority** while leveraging AI for efficiency 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 real-world 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.

Moreover, 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.

Finally, we focus on a single-instructor workflow. Many educational settings, such as larger classrooms or institutional contexts, involve multiple graders, and *VeriGrader* partially supports this by incorporating exemplars from multiple instructors into the LLM prompt as a list, allowing each instructor's feedback to inform grading. However, when instructors offer conflicting judgments or feedback, the system currently does not fully resolve these discrepancies, which may affect fairness. Developing mechanisms for reconciling conflicting feedback and managing consistency represents an important direction for future work.

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 LLMs for subsequent grading through in-context learning. We present a usage scenario to demonstrate *VeriGrader*, and conduct a comprehensive user study with twelve participants to confirm the system's effectiveness. We also discussed observations and findings from the research process, which provide valuable insight into AI plus Education.

Acknowledgments

We sincerely thank all reviewers for their constructive comments. The work was supported by the Guangdong Basic and Applied Basic Research Foundation (2025A1515010162), the National Natural Science Foundation of China (62402184), the Guangdong Provincial Fund for Basic and Applied Basic Research—Regional Joint Fund Project (Key Project) (2023B1515120078), the Guangdong Provincial Natural Science Foundation for Outstanding Youth Team Project (2024B1515040010), the Science and Technology Planning Project of Guangdong Province (2025B0101120003), and the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN-2020-03966).

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*. doi:10.1145/3746059.3747670

[6] Roman Chinoracky and Natalia Stalmasekova. 2025. Ethical Problems in the Use of Artificial Intelligence by University Educators. *Education Sciences* 15, 10 (2025). doi:10.3390/educsci15101322

[7] Domenic V Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment* 6, 4 (1994), 284. doi:10.1037/1040-3590.6.4.284

[8] 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

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

[10] 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

[11] 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. doi:10.1145/3185517

[12] 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

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

[14] Kilem Gwet. 2001. Handbook of inter-rater reliability. *Gaithersburg, MD: STATAXIS Publishing Company* (2001), 223–246.

[15] Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *Brit. J. Math. Statist. Psych.* 61, 1 (2008), 29–48. doi:10.1348/000711006x126600

[16] 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

[17] 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

[18] 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. doi:10.1007/s40708-016-0042-6

[19] 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. doi:10.1177/01466216231174566

[20] 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

[21] 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

[22] 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. doi:10.3389/frai.2024.1456486

[23] 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

[24] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* (1977), 159–174. doi:10.2307/2529310

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

[26] Zilin Li, Weihan Zhang, and Jun Tao. 2025. FlowLLM: Large language model driven flow visualization. *Vis. Informatics* 9, 3 (2025), 100241. doi:10.1016/J.VISINF.2025.100241

[27] 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

[28] 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. doi:10.1177/0265532214530699

[29] 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. doi:10.1002/fer3.10

[30] 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

[31] 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

[32] 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

[33] 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

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

[35] 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. doi:10.18653/v1/2025.findings-acl.950

[36] Fazil T Najafi, Vani Ruchika Pabba, Rajarajan Subramanian, and Sofia M Vidalis. 2025. AI-Assisted Grading–A Study on Efficiency and Fairness. In *2025 ASEE Southeast Conference*. doi:10.18260/1-2--54139

[37] 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. doi:10.1186/s12909-024-05881-6

[38] Lele Sha, Mladen Rakovic, Alexander Whitelock-Wainwright, David Carroll, Victoria M. Yew, Dragan Gasevic, and Guanliang Chen. 2021. Assessing Algorithmic Fairness in Automatic Classifiers of Educational Forum Posts. In *Artificial Intelligence in Education*. Springer International Publishing, Cham, 381–394. doi:10.1007/978-3-030-78292-4_31

[39] 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

[40] Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin* 86, 2 (1979), 420.

[41] 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

[42] Praveen Soni, Cyril de Runz, Fatma Bouali, and Gilles Venturini. 2024. A survey on automatic dashboard recommendation systems. *Vis. Informatics* 8, 1 (2024), 67–79. doi:10.1016/J.VISINF.2024.01.002

[43] 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. doi:10.1187/cbe.11-11-0100

[44] 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

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

[46] 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

[47] 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. doi:10.1145/3728963

[48] 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.

[49] 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

[50] 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

[51] 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 International Conference on Distance Education and Learning*. Association for Computing Machinery, 308–314. doi:10.1145/3675812.3675837

[52] 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

[53] 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

[54] 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

[55] Jiayi Zhou, Renzhong Li, Junxiu Tang, Tan Tang, Haotian Li, Weiwei Cui, and Yingcai Wu. 2024. Understanding Nonlinear Collaboration between Human and AI Agents: A Co-design Framework for Creative Design. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 170:1–170:16. doi:10.1145/3613904.3642812

[56] 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.