EvaluAId: Human-AI Collaborative Evaluation of Open-Ended Student Essays

Chao Zhang*
Cornell University
Ithaca, New York, USA
cz468@cornell.edu

Kexin Phyllis Ju*
University of Michigan
Ann Arbor, Michigan, USA
kexinju@umich.edu

Xinyi Lu
Computer Science and Engineering
University of Michigan
Ann Arbor, Michigan, USA
lwlxy@umich.edu

Yu-Chun Grace Yen
Computer Science
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
yyen@nycu.edu.tw

Jeffrey M. Rzeszotarski
Department of Computer Science
Loyola University Maryland
Baltimore, Maryland, USA
jeff.rzeszotarski@gmail.com

Abstract

Open-ended writing assignments are central to higher education, yet heterogeneous submissions and scale make evaluation difficult. Automated writing evaluation (AWE) promises speed but often trades away transparency and sidelines human judgment. This paper repositions the AI as an on-demand collaborator that can provide specific, targeted support. In a formative study, we expose leverage points in three cognitive dimensions: evidence identification, comparative judgment, and feedback composition. Guided by these insights, we build EvaluAId, which supports interactive rubric-content mapping, adaptive benchmarking and self-calibration, and personalized, rubric-aligned feedback synthesis. Through a within-subjects study with 12 TAs, we evaluate how this approach supports grading compared with a rubric+LLM chatbot and an LLM-based AWE; EvaluAId improved alignment with expert ratings and increased graders' satisfaction. Finally, interviews with TAs, instructors, and students underscored the value of thoughtfulness supported by EvaluAId while surfacing practical considerations for integration into classroom. Together, our results argue for deliberate, evidence-first, human-in-the-loop evaluation.

CCS Concepts

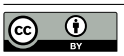
• **Human-centered computing** → **Interactive systems and tools**; **Empirical studies in HCI**.

Keywords

Writing evaluation, student essays, human-AI collaboration

ACM Reference Format:
Chao Zhang, Kexin Phyllis Ju, Xinyi Lu, Yu-Chun Grace Yen, and Jeffrey M. Rzeszotarski. 2026. EvaluAId: Human-AI Collaborative Evaluation of Open-Ended Student Essays. 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, 20 pages. https://doi.org/10.1145/3772318.3790814

*The first two authors contributed equally to this work.



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790814

1 Introduction

In a university, open-ended writing assignments require students to integrate domain knowledge, engage in higher-order reasoning, and produce original, well-structured arguments. Within User-Centered Design classes, for example, students may critique and compare designs using usability principles. Open-ended tasks have no single correct answer. Instead, students employ diverse strategies such as describing user scenarios, evaluating features sequentially, or drawing from personal experience. Such diversity is pedagogically valuable but challenging to assess rigorously.

To support evaluation of student essays, instructors often create **analytic rubrics** [6, 18, 39, 85, 86]: predefined grading dimensions with performance levels that aid both scoring and feedback. In the above example, a rubric might assess whether students address key usability issues and offer meaningful comparisons on a four-level scale (1 as not at all, and 4 as comprehensive). Graders use the rubric to assign scores and write feedback that justifies their evaluation and supports student improvement. Despite the use of rubrics, assessing open-ended essays remains challenging. Students often write in free-flow structures. Criteria may appear in different orders, and a single criterion can be scattered across the essay. Graders' standards may shift as more assignments are reviewed. They must also provide personalized feedback to create pedagogical value. As class size increases, grading essay assignments tends toward inconsistency and generic feedback, potentially compromising the assignment's worth.

Recent work in Human-Computer Interaction (HCI) and AI has explored end-to-end automated writing evaluation (AWE) systems to support large-scale assignment assessment and feedback generation [29, 51, 57, 67, 81, 93]. While such systems can reduce grading time for generic tasks (e.g., argumentative essays in standardized tests), they often fall short in classroom settings with open-ended, task-specific assignments in three key ways: (1) many systems rely on fixed, pre-defined rubrics and struggle to adapt to course- or instructor-defined criteria [88, 92]; (2) they prioritize surface-level qualities (e.g., grammar, coherence) [22, 66, 83, 96] over the

domain-specific competencies that open-ended assignments aim to assess; and (3) their one-shot, black-box judgments make it hard to trace decisions back to evidence, limiting graders' agency and the effectiveness of feedback [35].

In this paper, we reposition the AI not as an autonomous grader, but as an **on-demand collaborator that can provide specific, targeted support**. Our aim is to utilize the AI to augment specific, cumbersome tasks (e.g., identifying evidences across many assignments given a criterion) without permitting users to defer grading judgments to the system.

Through a formative study, implementation of tool called EvaluAId, and two follow-up lab evaluations, we demonstrate this reframing. Working with 12 TAs, we first examined challenges and strategies across the grading workflow, exposing leverage points in three processes: evidence identification, comparative judgment, and feedback composition. Guided by these insights, we implemented EvaluAId, a human-AI collaborative system that integrates instructor-defined rubrics with AI support to help graders evaluate open-ended student essays. EvaluAId breaks criteria into guiding questions that can be clicked to highlight relevant student text. It lets graders set a baseline essay and then surfaces similarities, strengths, and weaknesses to support pairwise comparisons. Finally, it supports feedback composition: graders select student text and rubric snippets and steer the AI to generate retrieval-grounded drafts. Across EvaluAId, the AI suggestions are opt-in—graders invoke, steer, inspect, edit, and approve them.

To understand how this collaborative approach supports grading in practice, we conducted a quantitative within-subjects study with 12 TAs comparing EvaluAId against an interactive rubric with an LLM chatbot and an LLM-based AWE. We found that TAs using EvaluAId achieved better alignment with expert scores and greater consistency, produced more actionable feedback for student improvement, and reported higher satisfaction with their performance. We then used EvaluAId as a probe in interviews with 12 TAs, 6 instructors, and 6 students to explore qualitative, multi-stakeholder perspectives on this human-AI collaborative approach. Results showed that EvaluAId engaged TAs in a thoughtful and deliberate grading workflow, which was seen as important for preserving agency and maintaining accountability by instructors. Students appreciated the visible human effort in the process but raised concerns about priming and consistency. We conclude by discussing how to foster thoughtfulness in human-AI collaborative grading and considerations for scaling EvaluAId in classroom settings.

In summary, this paper presents the following contributions:

(1) A formative study that identifies design opportunities and leverage points across three cognitive processes in rubric-based grading: evidence identification, comparative judgment, and feedback composition;
(2) EvaluAId, a human-AI collaborative system that augments graders across these processes for rubric-based evaluation of open-ended student essays, positioning human judgment as primary and AI as an on-demand collaborator that provides specific, targeted support;
(3) Empirical findings from a within-subjects study demonstrating EvaluAId's effectiveness in improving TAs' grading performance, compared to an interactive rubric+LLM chatbot and an LLM-based AWE;
(4) Perspectives from semi-structured interviews with TAs, instructors, and students that underscore the benefits and challenges of this human-AI collaborative approach in grading.

2 Related Work

2.1 Cognitive Processes in Rubric-Based Grading

Rubric-based grading demands complex cognitive labor [52]. Graders must interpret student writing, align it with rubrics, and make evaluative decisions (including scores and feedback) that are both consistent and justifiable [30]. Drawing from assessment literature and educational psychology, our work examines three core reasoning processes in the rubric-based evaluation: evidence identification, comparative judgment, and feedback composition.

Evidence identification concerns locating textual elements that correspond to specific rubric criteria [20]. This requires extracting and interpreting content that supports (or challenges) a given assessment [21, 45], yet the heterogeneity of student essays—diverse styles, implicit reasoning, and ambiguous language—places high demands on graders' interpretive skills. After identifying evidence, graders engage in comparative judgment: implicitly or explicitly comparing a submission against other work to calibrate their standards [55, 65]. Prior work suggests that pairwise comparisons can improve consistency and reduce rubric misalignment in subjective evaluations [9, 17]. However, maintaining a coherent mental representation of multiple pairwise comparisons and benchmarks is cognitively demanding. Beyond assigning scores, graders need to translate evaluations into formative feedback that links performance to improvement [25, 33, 76]. This synthesis blends textual evidence, evaluative reasoning, and rubric language into clear, constructive guidance [70]. Under time constraints and at scale, feedback composition can become one of the most cognitively challenging parts of grading. To understand how these mechanisms manifest in everyday practice and where support is most needed, we conducted a formative study with graders (i.e., student TAs). The study surfaced frictions, workarounds, and design opportunities within each process, directly informing the interaction design of our tool.

2.2 Automated Writing Evaluation

One common strategy for reducing the burden of assessment is automated writing evaluation (AWE) systems. They inspect longer-form student writing, typically producing holistic scores or targeted feedback via end-to-end machine learning pipelines [29, 51, 57, 67, 81, 93]. Such systems are widely used in universities and large-scale standardized assessments (e.g., GRE, TOEFL) [5, 24, 61, 73, 84]. However, most are trained on domain-specific corpora and assume fixed, general-purpose rubrics [88, 92], limiting adaptability to instructor-defined criteria and open-ended prompts. Some prioritize surface-level qualities (e.g., grammar, coherence) [22, 66, 83, 96] over the domain-specific competencies that open-ended assignments aim to assess. As a result, they fare better in constrained scenarios (e.g.,

argumentative writing in standardized tests [61]) than in diverse genres and authentic classroom contexts.

Recent LLMs have broadened access to AWE capabilities, as they can generate fluent, context-aware feedback [2, 7, 72, 74]. In some studies, they achieve scoring performance on par with or exceeding human raters [11, 36, 43, 49, 64, 88, 89], and they naturally further reduce the time necessary for assessment.

Despite these recent advances, the use of LLMs in assessment has not resolved long-standing concerns around automation, inaccuracies, and alignment [13, 31, 35, 41, 71]. First, end-to-end AI graders typically *automating* scoring or feedback without allowing graders to interrogate or contextualize the outputs, reducing educators' roles to passive overseers rather than active evaluators [14, 28]. This lack of transparency makes it difficult for instructors to trace decisions back to specific evidence in student writing [27]. Compounding this issue, LLM-based systems may produce fluent but inaccurate or fabricated feedback—so-called "hallucinations [1, 37, 48, 59]," which can mislead both graders and students if left unchecked. When all of these issues combine, such as when feedback that contradicts rubric intent or overlooks important content, there is limited recourse. In this work, we instead ask how might the AI *scaffolds* the processes that human graders enact during rubric-based evaluation. A human–AI teaming approach can reposition the AI as a collaborative facilitator, prioritizing educator agency.

2.3 Human-AI Collaboration in Educational Contexts

The HCI and learning sciences communities increasingly advocate collaboration over automation. Rather than delegating full control to AI, systems are being designed to support human-AI collaboration. In this line of work, prior studies have explored human–AI approaches across a wide range of contexts for supporting teaching [4, 50]. These efforts include visualizing student progress and learning challenges [54, 56], assisting with student team formation [90], improving classroom orchestration [53, 91], and supporting lesson and quiz preparation [16, 50].

In the domain of writing evaluation, recent work has focused on helping students make use of feedback. For example, Zhang et al. designed Friction to support students in reflecting on feedback they receive on their writing [94], and Synthia to help students interpret feedback and iteratively revise their work [95]. However, collaboration between human graders and AI systems to support effective writing evaluation remains relatively underexplored. One most closely related work is by Xiao et al. [88], who introduced a dual-process scoring system that combines end-to-end score generation with automated rationales, but their approach relies on fixed rubrics and domain-tuned corpora, limiting applicability to open-ended, instructor-defined settings. In contrast, we embed the AI within graders' reasoning processes, refrain from direct score prediction, and adapt to instructor-defined rubrics.

3 Study 1: Probing Essay-Evaluation Challenges and Strategies to Derive Design Goals

In university classes, student teaching assistants (TAs) are typically responsible for grading. We conducted semi-structured interviews with 12 student TAs to understand their natural workflows of grading open-ended writing assignments. We focused on the existing challenges TAs have in this task and their strategies of using rubrics. We then specifically examined areas where they may be leverage points for human-AI collaboration, synthesizing them into design goals. We aim to address the following questions:

RQ1.1 What challenges do TAs face when grading open-ended writing assignments?

RQ1.2 What strategies do TAs use to address these challenges, including how they engage with rubrics?

3.1 Methodology

3.1.1 Participants. 12 TAs from four universities participated in the study (8 female, 4 male); three were undergraduates and nine were graduate students. They reported between one and four semesters of TA experience across courses enrolling 20-200 students, all of which involved grading open-ended writing with analytic rubrics. Their grading responsibilities ranged from 2 to 18 assignments per semester, and typically involved 5-25 essays per assignment, with some handling as many as 50-100 submissions in total. Disciplines included Information Science, Computer Science, Cognitive Science, Digital Media, Education, Linguistics, and Communication. Each participant received a $15 gift card.

3.1.2 Study Procedure. We conducted remote, semi-structured interviews via Zoom (40-75 minutes), recorded with participant consent. Sessions began with background questions to contextualize each TA's experience and role, followed by an in-depth walkthrough of a recent grading session focusing on rubric use (e.g., "*How do you align rubric criteria with student performance to assign scores?*"), feedback composition (e.g., "*How do you formulate feedback for individual students?*"), and consistency management (e.g., "*How do you maintain consistency across multiple submissions?*"). When feasible and appropriately anonymized following IRB guidelines, participants shared screens to demonstrate their workflow, allowing observation of how they applied rubrics, wrote comments, and navigated challenges. We then elicited visions for an ideal workflow and probed ethical boundaries for AI integration, including attitudes toward potential AI roles (e.g., assigning scores, drafting feedback) and limits on involvement (e.g., "*Where should AI's involvement end?*"). Interviews were transcribed verbatim and analyzed using reflexive thematic analysis [8].

3.2 Findings

Table 1 summarizes key grading dimensions from our TA interviews, with each dimension including challenges labeled as **C** and strategies as **S**. The following section elaborates on each dimension and provides illustrative quotes.

3.2.1 Locating Rubric Evidence Is Labor-Intensive. TAs must locate specific content within student submissions to justify scores against rubric criteria. However, variation in length, structure, and writing style across submissions makes this difficult **[C1]**. For instance, key arguments may be buried in verbose essays. As P2 noted, "*There is always something that is not relevant... I am frustrated about distinguishing them (from the content I need).*" With rubrics containing

Table 1: The summary of challenges and strategies reported in §3.2

Dimension	RQ1: Challenges	RQ2: Strategies
Evidence Identification	**C1**. Heterogeneous submissions make evidence hard to locate.	**S1**. Display rubric and essay side by side to keep criteria visible.
	C2. High volume and repetition cause fatigue and inconsistency.	**S2**. Manually tag rubric-relevant content to aid later judgment
Comparative Judgment	**C3**. Subjectivity and standard drift over time with fatigue.	**S3**. Establish adaptive "baseline" submissions as anchors.
	C4. Tracking quality across essays is cognitively demanding.	**S4**. Anchor new scores via pairwise comparison with baselines.
Feedback Composition	**C5**. Composing feedback that justifies scores is time-consuming.	**S5**. Combine rubric descriptors with specific student excerpts.
	C6. Composing feedback for improvements is time-consuming.	**S6**. Reuse or adapt pre-written phrases under time pressure.
Use of AI in Grading	**C7**. AI-generated scores may be inaccurate or unaccountable.	**S7**. Use AI for supportive tasks, not for final decisions.
	C8. AI outputs may lack nuance for subjective criteria.	**S8**. Prefer transparency about AI contributions and provenance.
	C9. Risk of over-reliance degrading grading quality.	**S9**. Always apply human review and final judgment.

layered criteria and dozens of submissions to grade, the process becomes repetitive and mentally exhausting **[C2]**, sometimes leading to reduced care over time. To cope, TAs adopted makeshift workflows such as split-screen setups **[S1]** and manually highlighting or tagging relevant content **[S2]**. Despite these efforts, the labor remains high, motivating interest in features that could automate evidence highlighting.

3.2.2 TAs Rely on Self-Calibration to Stay Consistent. Grading open-ended assignments is inherently subjective and prone to drift over time, especially under fatigue **[C3]**. As P5 shared, "*After grading 20 essays, my standards drift...*" TAs also struggle to maintain stable interpretations of rubrics across diverse responses **[C4]**. To combat inconsistency, they develop self-calibration strategies centered on implicit, content-driven benchmarks **[S3]** and cross-submission comparisons **[S4]**. Many anchor their evaluations by identifying similar submissions to serve as internal reference points, though these baselines remain fluid as new interpretations of the prompt emerge. As P12 explained, "*The baseline shifts when students approach the same prompt in wildly different ways.*" With a baseline in mind, TAs commonly make mental pairwise comparisons (e.g., "*Does the next one meet, exceed, or fall short of this standard?*"from P2). These practices point to the need for tools that support adaptive benchmarking, track and compare past decisions, and assist in rubric-aligned pairwise assessments.

3.2.3 Personalized Feedback Requires Time-Consuming Synthesis. TAs emphasized the pedagogical value of tailored feedback but noted that composing comments that both justify scores **[C5]** and guide improvement **[C6]** is one of the most time-consuming parts of grading. Many resort to copying and pasting rubric descriptions due to time pressure. Effective feedback requires mapping rubric levels to specific student content, yet vague language (e.g., "*clear explanation*") complicates alignment. As P7 remarked, "*Definitions of 'clear' differ.*" To cope, TAs manually combine rubric phrasing with contextually relevant student content **[S5]** and reuse or adapt pre-written phrases across submissions to save time **[S6]**. These efforts call for intelligent scaffolds that streamline personalized feedback synthesis.

3.2.4 TAs Demand Control and Accountability in AI Grading. TAs expressed concern about AI autonomy in grading, especially with regards to accountability if scores are inaccurate **[C7]** or if the AI lacks nuance for subjective criteria **[C8]**. "*Who takes responsibility if the AI grades it wrong?*" asked P3. Over-reliance on the AI could also erode grading quality **[C9]**. That said, many welcomed AI assistance for locating relevant content or drafting feedback **[S7]**, provided systems are transparent about AI contributions **[S8]** and they retain final decision-making authority **[S9]**. As P4 emphasized, "*The essay must be read by people... AI can help, but humans must decide.*" These views underscore the importance of preserving human agency and ensuring clear accountability in AI-augmented assessment workflows.

3.3 Design Goals

Building on prior work on cognitive processes in rubric-based grading (§2.1) and insights from our formative study, we translate these theoretical foundations and empirical findings into four design goals (**DGs**) that guide our system design presented in the next section. Guided by Gregor et al.'s schema for design principles, i.e., specifying aims, mechanisms, contexts, and rationales [23], we present these goals as theoretically informed, prescriptive statements intended to inform future work. Table 2 summarizes the details of each goal.

- **DG1 Interactive Rubric-Evidence Mapping.** To reduce the cognitive load of locating and justifying rubric-aligned evidence **[C1-C2]**, this design builds on assessment and professional noticing theories [20, 21, 45, 52] by decomposing each rubric criterion into concrete considerations and highlighting candidate textual spans for each consideration for grader review.
- **DG2 Adaptive Benchmarking Through Pairwise Comparison.** Drawing on comparative judgment research [9, 17, 55, 65] and formative observations that TAs self-calibrate through informal comparisons **[C3-C4, S3-S4]**, this design supports pairwise comparisons against grader-chosen baselines by leveraging AI to identify comparable submissions and surface relative strengths and weaknesses across essays.
- **DG3 Scaffolded Evidence-Grounded Feedback Synthesis.** To help TAs compose personalized, improvement-oriented feedback under time pressure, this design draws on formative feedback theory [33, 70] and formative findings on TAs' coping strategies **[C5-C6, S5-S6]** to scaffold comment writing by synthesizing rubric descriptors and selected evidence from student essays into editable AI-generated drafts.
- **DG4 AI as an On-Demand Collaborator that Keeps Humans in the Loop.** Informed by human-AI interaction guidelines [3,

Table 2: Design goals for human-AI collaborative grading, derived from prior research on cognitive processes in rubric-based evaluation (§2.1) and insights from our formative study (§3.2).

Design Goals	Aim	Context	Mechanism	Rationale
DG1. Interactive Rubric-Evidence Mapping	Help TAs efficiently locate and justify rubric-aligned evidence while preserving interpretive authority.	Grading open-ended, heterogeneous student writing requires TAs to locate rubric-aligned evidence across lengthy, stylistically diverse submissions, often causing fatigue and inconsistency [C1-C2].	Provide interactive AI-assisted mappings that decompose rubric criteria into concrete considerations, highlight candidate textual spans, and let graders toggle, confirm, and link highlights to rubric levels.	Grounded in assessment and professional noticing theory [20, 21, 45, 52], and supported by formative findings [S1-S2], this interaction scaffolds rubric interpretation and evidence identification.
DG2. Adaptive Benchmarking Through Pairwise Comparison	Support graders in maintaining consistent and calibrated judgments across subjective assessments.	Graders' standards drift across batches of subjective writing and rubric interpretations evolve over time [C3-C4].	Enable graders to set baseline essays, leverage AI to identify comparable submissions, and highlight relative strengths and weaknesses between paired essays to support self-calibration.	Comparative judgment research [9, 17, 55, 65] and formative evidence [S3-S4] show that structured comparisons formalize existing mental benchmarks, improving consistency.
DG3. Scaffolded Evidence-Grounded Feedback Synthesis	Enable TAs to compose personalized, improvement-oriented feedback efficiently.	TAs value personalized formative feedback but face time pressure, vague rubric language, and reuse generic comments [C5-C6].	Provide editors that bind feedback units to selected evidence and rubric levels and generate editable drafts referencing these links for both justification and improvements.	Building on feedback theory [33, 70] and observed coping strategies [S5-S6], scaffolding the synthesis of evidence and rubric information during feedback composition helps TAs produce specific, improvement-oriented comments.
DG4. AI as an On-Demand Collaborator that Keeps Humans in the Loop	Ensure accountability and trust in AI-assisted grading by preserving human control.	TAs express concern over AI autonomy, accountability, and grading errors in high-stakes contexts [C7-C9].	Keep AI advisory and transparent: label AI contributions, require human confirmation, and ensure all AI highlights and outputs remain editable by graders.	Consistent with prior design guidelines for human-AI interaction [3, 12, 69] and our findings [S7-S9], ensuring that the AI functions as an on-demand collaborator whose support is traceable and overrideable, rather than as a final decision-maker, preserves trust and accountability in AI-assisted assessment.

12, 69] and formative findings **[C7-C9, S7-S9]**, this design positions AI as an on-demand collaborator that provides targeted support while ensuring its contributions remain transparent, traceable, and fully editable by graders to preserve human control and accountability.

4 The EvaluAId System

EvaluAId is a web-based system that integrates instructor-defined rubrics with the AI support to guide graders in essay evaluation and personalized feedback. The interface features a Grading Canvas for organizing graded and ungraded essays (Fig. 1A), a Rubric Workbench for navigating criteria and quality levels (Fig. 1B), an Essay Reader (Fig. 1C), and a Feedback Composer (Fig. 3). Graders begin by uploading student essays with a rubric. They can also add lecture slides, readings, or exemplar responses to give the system more context and align grading standards with expectations (Fig. 1D). While interactive rubric development is a compelling area for future work, in EvaluAId we focus on evaluation with existing, well-structured rubrics.

Throughout EvaluAId, the AI acts as an **on-demand assistant**: it surfaces suggestions, highlights, and comparisons that graders explicitly invoke, inspect, edit, and approve. Nothing is auto-scored. Provenance (e.g., selected snippets and retrieved sources) is kept visible. Similarity meters and visual encodings are **advisory, not prescriptive**. This promotes **human agency and accountability [DG4]** while letting the AI reduce search and drafting overhead. These components work together to streamline evidence identification **[DG1]**, support benchmarking and self-calibration **[DG2]**, and scaffold personalized feedback synthesis **[DG3]**, as detailed in the following subsections.

4.1 Automatic Rubric–Content Mapping [C1–C2 → DG1; informed by S1–S2]

To address graders' difficulty locating rubric-aligned evidence in heterogeneous submissions, EvaluAId automatically decomposes each rubric criterion into concise, clickable considerations (Fig. 1E). In Rubric Workbench, graders can switch criteria via tabs and toggle considerations to control what the system highlights, which immediately color-marks matched spans in the essay (Fig. 1F) so attention goes to relevant passages as skimming the full-text.

A vertical distribution mini-map in Essay Reader (Fig. 1G) visualizes the relevant-text distribution for each criterion, revealing where evidence is concentrated or sparse across the assignment and aiding navigation. Graders can click rectangular cells in this map to switch the active criterion (and its considerations), which in turn updates highlights in Essay Reader. This addresses attention fatigue **[C1-C2]** and mapping **[DG1]** while keeping graders in control of which considerations are active, how evidence is interpreted, and when any mapping informs scoring **[DG4]**.

As shown in Fig. 2, to locate relevant content from student essays for a given criterion in a rubric, we first prompt `GPT-4o` to break down long descriptors into several consideration questions. These

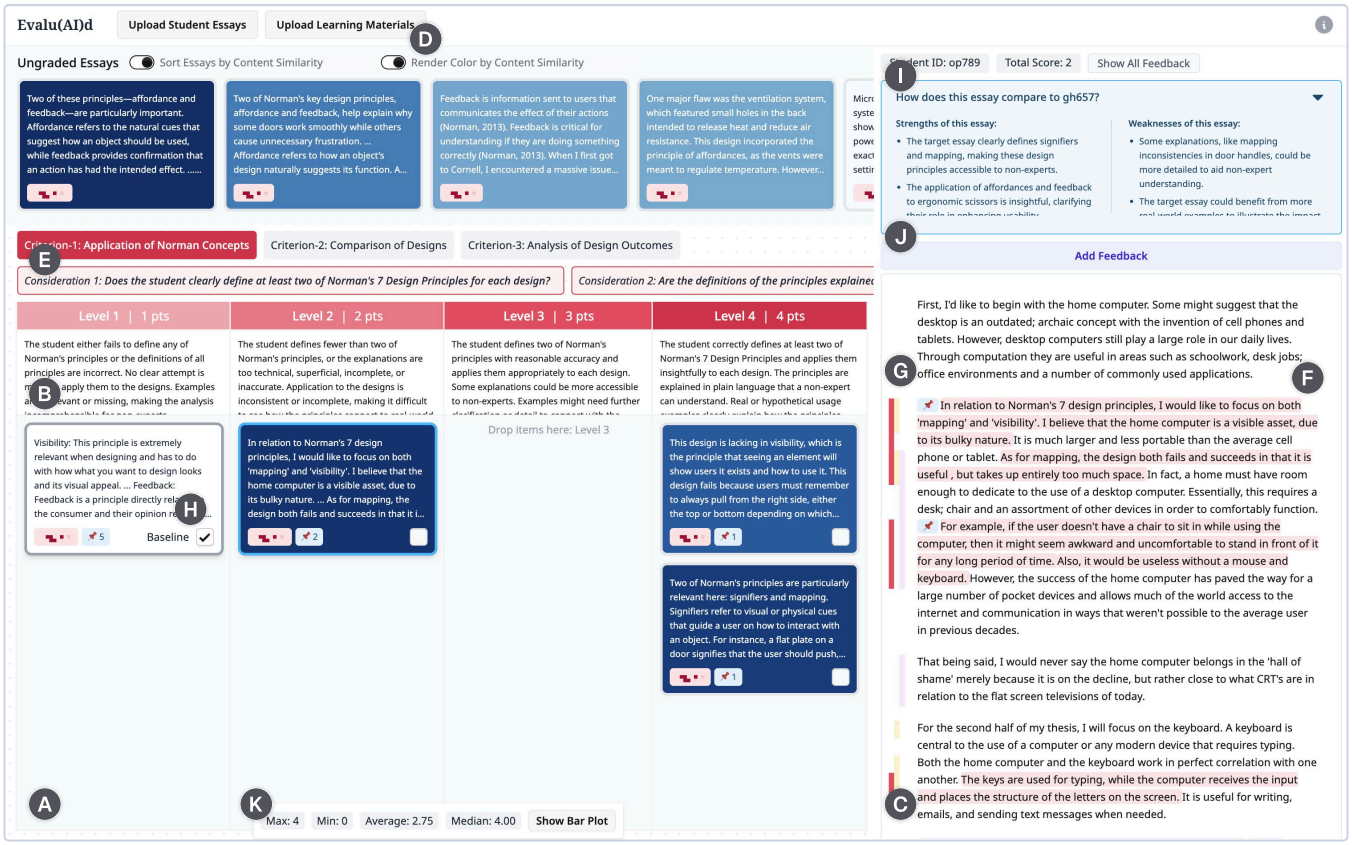


Figure 1: EvaluAId's UI: (A) Grading Canvas organizes graded and ungraded submissions, with options to sort or color by similarity to a baseline (H); (B) Rubric Workbench displays criteria and quality levels which can be toggled (E) to drive evidence highlighting while a statistics view (K) monitors calibration; and (C) Essay Reader permits consideration-activated highlights (F), with a side evidence map (G) summarizing the distribution of relevant text, viewing strengths and weaknesses relative to the baseline (I), and launching the Feedback Prep Station via Add Feedback (J; see Fig. 3). The upload bar (D) manages materials.

questions represent what students should address when completing the essay. We then pair each consideration question with each student essay and prompt `GPT-4o` to identify sentences from the given essay that directly answer or relate to the question. We will later evaluate the performance of this LLM pipeline in §4.5.1.

4.2 Adaptive Benchmarking and Self-Calibration [C3–C4 → DG2; consistent with S3–S4]

To stabilize standards over long grading sessions, graders pin a baseline essay that serves as an anchor for comparative judgment, and they can change this baseline at any time to reflect new reference points as their understanding of the cohort evolves (Fig. 1H). This aligns with typical practices where graders locate touchstone examples for use as comparison cases. EvaluAId computes content similarity from the current baseline to all remaining submissions and encodes it with a shared hue whose depth indicates similarity; graders can sort and reorder ungraded essays accordingly **[S3–S4]**. To measure content similarity, we first convert the relevant content for each criterion in each student essay into embeddings using OpenAI's `text-embedding-3-large` model. Then, we compute the cosine distance between the embedding from baseline essay and the embeddings from other essays to obtain the similarity scores. This approach follows established practices in semantic textual similarity, where embeddings have been shown to effectively capture nuanced semantic relationships in text [47, 77].

For any open essay, Essay Reader surfaces strengths and weaknesses relative to the active baseline, providing a fast calibration cue before detailed scoring (Fig. 1I). The drag-and-drop Grading Canvas (Fig. 1A) and a statistics view (Fig. 1K) further externalize the grader's evolving mental picture of the cohort, reducing memory burden when tracking many submissions **[C4]** and fulfilling **[DG2]**. Critically, graders choose and adjust the anchor, control the ordering, and treat similarity cues as guidance to be judged and not as final decisions **[DG4]**.

4.3 Personalized and Rubric-Aligned Feedback Synthesis [C5–C6, C7–C9 → DG3; grounded in S5–S6]

When composing feedback, graders click "Add Feedback" (Fig. 1J) to open Feedback Composer and select essay snippets (student text) and rubric descriptor snippets (the relevant consideration or

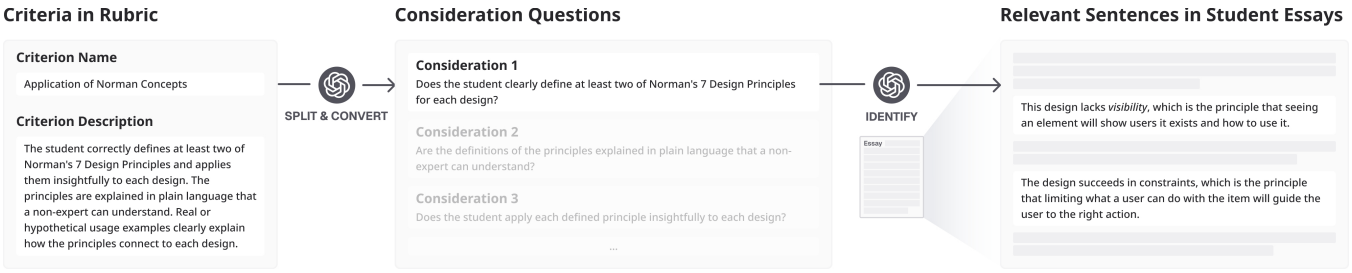


Figure 2: EvaluAId's LLM pipeline in locating relevant content from student essays for a given criterion.

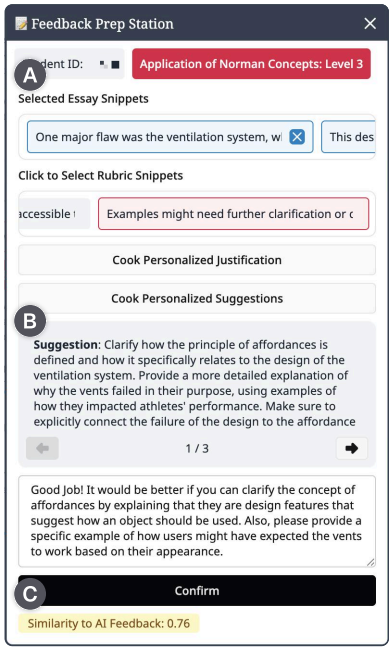


Figure 3: EvaluAId's Feedback Composer, used for writing personalized, rubric-aligned feedback. (A) Graders select essay and rubric snippets to condition retrieval-augmented generation. (B) AI proposes editable draft justifications or suggestions based on the selected content and uploaded learning materials; graders can browse multiple alternatives and revise freely. (C) A similarity-to-AI indicator shows how closely the edited feedback matches the model's draft, and the Confirm button records only human-approved feedback.

level language) (Fig. 3A). These selections prompt the AI to draw on the uploaded learning materials to propose grounded justifications and actionable suggestions tied to the chosen evidence. When clicking "Cook", the system prompts GPT-4o to synthesize rubric snippets and student essay excerpts provided by users into a draft justification or suggestion. Drafts are presented as editable alternatives (Fig. 3B): graders can revise wording, mix candidates, or start from scratch (addressing **[C5-C6]** and realizing **[DG3]**). The performance of our LLM-based feedback synthesis method is evaluated in §4.5.2.

A passive similarity-to-AI text highlight offers transparency about how closely the final text tracks the model's draft, but the confirm step ensures only human-approved feedback is recorded (Fig. 3C). Once confirmed, each essay card displays a corner tag indicating the number of inline feedback comments for this student. By making selection, grounding, editing, and approval explicit, this workflow preserves human authority and controllability while leveraging the AI for retrieval and first-pass drafting **[DG4]**.

4.4 Implementation Notes

EvaluAId is implemented through `Next.js`, which supports server-side rendering for efficient API calls including requests to the `OpenAI APIs` for instructing pre-trained GPT models, and the `Firebase` APIs for logging user events. Retrieval-augmented generation functionality is built with `Vercel AI`, `Drizzle ORM`, and `Neon Database`, enabling fast retrieval from the uploaded learning materials based on embeddings during feedback generation. The drag-and-drop feature is implemented with `dnd-kit`, while `D3.js` supports interactive visualizations such as rubric-content mapping and similarity color encodings.

We prompt `GPT-4o` to (i) locate relevant content from student essays according to active considerations (detailed in §4.1; evaluated in §4.5.1), (ii) generate comparative summaries (strengths and weaknesses) between the current and baseline essays (used in §4.2), and (iii) produce feedback grounded in the specified rubric descriptors and essay snippets (detailed in §4.3; evaluated in §4.5.2). Prompts and example outputs are provided in Supplementary Materials.

4.5 Technical Evaluation

We validate the technical capabilities of our LLM pipelines implemented in EvaluAId, acknowledging that LLMs may be prone to hallucinations or other inaccuracies [37] which risk misguiding users or diminishing the overall usability of the system. Specifically, we evaluate whether the LLM pipelines can (1) accurately locate relevant content from the student responses according to criteria considerations and (2) generate high-quality, customized, and accurate feedback (i.e., justifications and suggestions).

4.5.1 Evaluating Performance of LLM Pipelines in Content Location. There is no established dataset for content location based on scoring criteria, but the process of finding content based on consideration questions shares the essence of a canonical task: machine reading comprehension (MRC).

Dataset: In the MRC field, the second version of the Stanford Question Answering Dataset (SQuAD 2.0) is the most widely used benchmark for evaluating model performance. It covers a wide range of topics from Wikipedia. However, SQuAD 2.0 is designed for general purposes rather than educational use, and the located content is typically short phrases rather than full sentences. To address this, we adapted SQuAD 2.0 by expanding each located phrase into the full sentence in which it appears. Additionally, we included another educational-purpose dataset, the Student Essay Dataset (SED), which contains pairs of task fulfillment queries and the locations of relevant sentences in student essays. However, the length of student essays in SED is shorter than the Wikipedia articles in SQuAD 2.0. To leverage the strengths of both datasets, we sampled 120 Q&A pairs from each. The average word count of the context is 386 for SQuAD 2.0 samples and 150 for SED samples.

Results: We calculated ROUGE-1 recall (the proportion of individual words in the reference text that also appear in the generated text), precision (the proportion of individual words in the generated text that also appear in the reference text), and F1 scores. In the SQuAD 2.0 samples, our LLM pipeline achieves 0.91 in recall, 0.92 in precision, and 0.89 in F1, outperforming the baseline BERT model, which achieved 0.73 in F1. Similarly, in the SED samples, our LLM pipeline achieves 0.83 in recall, 0.86 in precision, and 0.81 in F1, also surpassing the baseline BERT model, which achieved 0.68 in F1. These results demonstrate the effectiveness of our LLM pipeline in locating relevant content based on queries, outperforming the baseline model in both datasets.

4.5.2 Evaluating Performance of LLM Pipelines in Feedback Synthesis. The goal of our LLM-based feedback generation pipeline is to produce **high-quality**, **customized**, and **accurate** justifications and suggestions based on the given rubric and essay snippets. In this section, we computationally evaluate how well this goal is achieved.

Dataset: Since no established dataset exists for rubric-based, criterion-by-criterion assessment of constructed-response writing assignments, we utilized the Automated Student Assessment Prize Short Answer Scoring Dataset (ASAP-SAS). This dataset is well-suited to our study because: (1) it provides all the contextual information required for our prompt pipeline, including the task, rubric, student responses, and scores; (2) each rubric focuses on a single dimension related to task fulfillment, allowing the descriptions of each level to be included in our prompt as criterion snippets; and (3) the student responses are short, making them comparable to essay snippets relevant to the criterion snippets. For our evaluation, we sampled 100 data points from the dataset, covering topics such as English, science, and biology. We input them to our LLM pipelines and generated 100 justifications and 100 suggestions.

Results: We input the sampled data points to our LLM pipelines and generated 100 pairs of justifications (length: $Min = 31$, $Max = 65$, $M = 45.81$, $SD = 7.66$) and suggestions (length: $Min = 30$, $Max = 78$, $M = 49.06$, $SD = 8.22$).

- **Quality:** Following the computational linguistic methods proposed by Krause et al. [44] for predicting the helpfulness of feedback, we evaluate whether the generated justification is reasoned (i.e., includes explanatory reasoning) and whether the generated suggestion is actionable (i.e., contains commands, suggestions, or hypothetical situations). Additionally, prior work has suggested that negative feedback can threaten one's ego and reduce feedback effectiveness[10]. To assess whether the generated feedback is non-negative, we compute the sentiment scores for both justifications and suggestions on a scale ranging from -1 to 1, with 1 indicating the most positive sentiment. The results show that 91% of the generated justifications are reasoned, and 98% of the generated suggestions are actionable. The sentiment of both justifications ($Min = -0.25$, $Max = 0.68$, $M = 0.08$, $SD = 0.18$) and suggestions ($Min = -0.12$, $Max = 0.40$, $M = 0.10$, $SD = 0.14$) is overall neutral.
- **Customization:** In this context, feedback is considered customized if it is specific to the provided rubric snippets and personalized to the corresponding student content. Customization is quantified using two similarity metrics: (1) similarity to the given rubric snippets and (2) similarity to the given student content. To compute these metrics, we transformed the generated justifications, generated suggestions, rubric snippets, and student content into embeddings using OpenAI's text-embedding-3-small model[1] and measured the cosine similarity between the vectors. For comparison, we also calculated the similarity of each justification and suggestion to a random rubric and a random essay. Paired t-tests revealed that the similarity ($M = 0.61$, $SD = 0.13$) of the generated justifications to the given rubric was significantly higher than their similarity ($M = 0.42$, $SD = 0.06$) to a random rubric ($t(99) = 14.25, p < .001^{***}$). Similarly, the similarity ($M = 0.52$, $SD = 0.21$) of the generated suggestions to the given rubric was significantly higher than their similarity ($M = 0.28$, $SD = 0.06$) to a random rubric ($t(99) = 8.98, p < .001^{***}$). We also observed that the similarity ($M = 0.45$, $SD = 0.12$) of the generated justifications to the given student content was significantly higher than their similarity ($M = 0.34$, $SD = 0.09$) to a random essay ($t(99) = 12.58, p < .001^{***}$). Additionally, the similarity ($M = 0.54$, $SD = 0.12$) of the generated suggestions to the given student content was significantly higher than their similarity ($M = 0.45$, $SD = 0.10$) to a random essay ($t(99) = 5.59, p < .001^{***}$).
- **Accuracy:** LLMs are prone to mistakes and hallucinations which can mislead graders, hindering students' learning. In this context, hallucinations are defined as responses that contain fabricated content unfaithful to the materials in the task prompt or inaccurate facts, concepts, or numbers that conflict with validated knowledge. Two research assistants manually inspected all the generated justifications and suggestions from prior evaluations in this section ($IRR = 0.84$)). The results show that only 7% of the generated justifications and 6% of the generated suggestions were classified as hallucinations by at least one evaluator. This indicates a very low likelihood of hallucinations in the generated feedback, with

[1] https://platform.openai.com/docs/guides/embeddings

further resilience gained from a system design which discourages any direct use of machine output in student-facing materials.

In sum, our LLM-based method produces customized and accurate feedback drafts. It generates reasoned, actionable responses with neutral sentiment, aligns closely with rubric and student content, and exhibits a low error rate.

5 Study 2: Assessing Effectiveness and Perceived Support of EvaluAId in Grading Tasks

To further evaluate EvaluAId in grading tasks, we conducted a within-subjects study[2] with 12 student TAs who typically grade open-ended writing assignments to evaluate how EvaluAId assists with their grading workflow compared to a baseline system powered by a conversational AI assistant. In Study 3 we will revisit EvaluAId from the perspective of multiple stakeholders, but as in modern universities the majority of grading is performed by TAs, we felt that they were the most ecologically valid population for this evaluation. We aim to address the following research questions:

RQ 2.1 How effective is EvaluAId in supporting TAs with grading tasks?

RQ 2.2 How do TAs perceive the support provided by EvaluAId for grading tasks?

5.1 Methodology

5.1.1 Baseline. To contextualize the value of EvaluAId, we compared it against a baseline condition using a general-purpose LLM-powered conversational assistant (CA). This baseline reflects current grading tool design such as Canvas SpeedGrader[3] and the growing trend of integrating tools like ChatGPT into grading workflows [43, 64, 75]. The baseline tool featured an interface similar to EvaluAId but without dedicated support for evidence identification, comparative judgment, and feedback composition. As users navigated between essays and criteria, the system automatically updated the prompt with relevant content, ensuring smooth interactions and preventing usability issues such as manual copy-and-pasting.

Participants in the baseline could engage with the assistant to ask questions or request help with grading or feedback writing but were not required to use it in any particular way, mirroring realistic usage patterns in educational settings. This also helps to reduce potential confounds from priming participants to use the model to complete tasks they may otherwise prefer to do manually even in the presence of an assistant. The UI layout, rubric structure, and LLM model were kept consistent across both systems.

5.1.2 Task Materials. We developed two parallel sets of essay responses to use as task materials. Each set of essays responds to a classical critique assignment (and their associated rubrics) from an entry-level HCI course at a private university in the United States. One assignment asked students to apply Norman's Design Principles to evaluate physical designs, while the other focused on applying Nielsen's Usability Heuristics to virtual designs. Both assignments required students to analyze two given designs using the corresponding principles or heuristics and to support their evaluations with concrete examples of user interactions. This setup allowed us to introduce diversity in content while maintaining comparable complexity and cognitive demands across the two grading materials. Supplementary Materials provide the full writing prompts and rubrics.

Essays: For each assignment, we recruited eight university students from relevant fields (e.g., art, design, information science) to write short essays, as IRB and local regulations prohibited us from using student submissions from the original courses. Writers received compensation of one research credit or $8. The two sets of essay responses were comparable in word count, ranging from 400 to 500 words ($t(18) = .909, p = .379$). To establish a robust ground truth and ensure that the two essay sets were reasonably comparable in quality for participants in our study, two introductory HCI course TAs independently graded all essays using the same rubrics that would later be used by our participants ($ICC = 0.78$). Discrepancies were later resolved through discussions with professors. No significant difference in quality was observed between the two essay sets ($t(18) = -.372, p = .715$).

Rubrics: The rubrics used by both the aforementioned evaluators and the study participants were adapted from the original course grading criteria and lightly revised by an experienced, independent HCI professor to improve clarity. Each rubric consists of three criteria, with four performance levels ranging from 1 (low quality) to 4 (high quality), accompanied by descriptive benchmarks for each level. The three criteria are "Application of Principles," "Comparison of Designs," and "Analysis of Design Outcomes." The descriptive benchmarks for each criterion in the two rubrics vary slightly depending on the specific design principles involved. Due to time constraints, participants in the study evaluated essays based only on the first two criteria.

5.1.3 Participants. We recruited 12 TA participants (Table 3; T01—T12); 9 female, 3 male; ages ranging from 23 to 29; $M = 25.58$, $SD = 2.23$) from U.S. universities through flyers, word-of-mouth, and internal participant recruitment systems. The participants were required to (1) have at least one semester of TA experience, (2) have experience in grading writing assignments using rubrics, (3) have taken this course or similar alternatives, and (4) be familiar with the knowledge and concepts involved in the task materials. Each participant received a $20 gift card as compensation for their time.

5.1.4 Study Procedure. The study began with obtaining informed consent from each participant, followed by a demographic survey. Participants then completed two task sessions, each beginning with a 3–5-minute tutorial and followed by a grading task using either EvaluAId or the CA system. To control for order effects, participants used EvaluAId and the CA system with different sets of task materials (see §5.1.2), assigned in a counterbalanced order.

In each session, participants were given 30 minutes to assess all essays in the assigned dataset using the respective system. The 30-minute time limit was determined through multiple rounds of pilot testing to ensure participants had sufficient time to engage meaningfully with the grading task without fatigue. Participants were asked to approach the task as if they were grading in a real-world

[2] All studies in this paper received approval from our institution's IRB.
[3] https://community.canvaslms.com/t5/Canvas-Basics-Guide/What-is-SpeedGrader/ta-p/13

Table 3: Demographic information of university TAs in Study 2 and 3.

ID	Gender	Age	TA Experience	AI Grading	AI Usage
T01	Female	28	0-6 months	Not yet but interested	N/A
T02	Female	28	1-2 years	Yes, occasionally	ChatGPT for feedback rephrasing
T03	Female	24	0-6 months	Not yet but interested	N/A
T04	Female	23	1-2 years	Yes, occasionally	ChatGPT for summarization
T05	Female	23	6 months-1 year	Yes, occasionally	ChatGPT
T06	Female	24	0-6 months	Not yet but interested	N/A
T07	Female	29	6 months-1 year	Yes, occasionally	ChatGPT for feedback rephrasing
T08	Male	28	2-5 years	Yes, frequently	ChatGPT for feedback rephrasing
T09	Female	24	6 months-1 year	I tried it once or twice	ChatGPT for automated grading
T10	Male	25	1-2 years	Not yet but interested	N/A
T11	Female	27	0-6 months	Yes, frequently	Grammarly, ChatGPT for evidence searching
T12	Male	24	6 months-1 year	I tried it once or twice	Grammarly, ChatGPT

scenario, where their scores would be sent directly to students, emphasizing the importance of accountability in their decisions. Following each condition, participants completed a short survey evaluating their perceptions of the AI support.

After completing both main conditions, participants took part in a 25-minute semi-structured interview. They were first briefly exposed to a fully automated end-to-end writing evaluation (AWE) system. The system was implemented using GPT-4o, with prompts designed based on prior work on LLM-based AWE [11, 36, 72]. Participants were shown the system-generated scores for the same essays they had previously graded and given 5 minutes to review the outputs. This phase was intended to elicit participants' reactions to such fully automated assessment. After this brief review, participants completed the same perception survey for AWE. Finally, they reflected on their experiences with each system, shared their attitudes toward human-AI collaboration in grading, and offered suggestions for improvement[4]. All study sessions were conducted remotely via Zoom. The procedure of one session is outlined in Fig. 4.

5.1.5 Measures. The post-session survey included five items (i.e., Satisfaction, Think-through, Collaboration, Transparency, and Control) from Wu et al. [87] to assess participants' perceived AI experience, four additional items to measure perceptions of Trust, Reliability, Fairness, and Justifiability, and one item to assess participants' Confidence in delivering the results to students. All survey questions are presented in Supplementary Materials.

To objectively evaluate grading performance in each condition, we computed the agreement between participants' assigned scores and expert-established ground truth using Root Mean Square Error (RMSE) and Pearson's correlation coefficient (two commonly used metrics in prior work [32, 58, 89]). Additionally, we assessed inter-rater consistency among participants using the Intraclass Correlation Coefficient (*ICC*) and compared the standard deviations across conditions.

For feedback evaluation, two TAs from an introductory HCI course (who did not participate in the study) independently rated the quality of 48 written feedback entries (24 per system condition) without knowledge of the condition under which each was produced. Following the methods and criteria developed by Steiss et al. [72], the feedback was evaluated across four dimensions: Accuracy, Criteria-Based, Clear Directions for Improvement, and Supportive Tone, using a 5-point Likert scale. Significant rating discrepancies (greater than a 2-point difference) were resolved through discussion and re-evaluation.

[4]These interview data will be analyzed in the next study presented in §6.

5.2 Analysis and Findings

5.2.1 RQ2.1: TA Performance in Grading. To assess the general effectiveness of EvaluAId in supporting TA grading, we evaluated differences and correlations between participants' scores and expert ratings, inter-rater consistency among participants, and the feedback written by participants. We conducted paired t-tests to compare EvaluAId with the baseline.

Differences with Expert Scores: Averaged by criterion, the mean score from EvaluAId was 2.688 (SD = .892), and from the CA baseline was 2.719 (SD = .856), compared to expert ratings of 2.875 (SD = 1.100). We further evaluated differences between TA scores and expert scores using RMSE (Fig. 5A). EvaluAId achieved significantly higher grading accuracy than the baseline ($t(11) = -2.339, p = .039^{*}$): RMSE was lower for EvaluAId ($M = .744, SD = .200$) compared to the CA baseline ($M = 1.046, SD = .347$). As context, the mean score from AWE was 3.125 (SD = .976), with an RMSE of 1.031.

Correlation with Expert Scores: We evaluated trend alignment with expert ratings using Pearson correlation (Fig. 5B). EvaluAId achieved significantly higher Pearson correlation with expert scores ($M = .754, SD = .129$; $t(11) = 2.479, p = .031^{*}$) than the CA baseline ($M = .509, SD = .281$). AWE's correlation with expert scores was 0.526, indicating moderate agreement, consistent with findings from previous research [43].

Inter-Rater Consistency: To compare inter-rater consistency, we calculated *ICC*. Results showed that participants using EvaluAId were more consistent (ICC = .612) than using the CA baseline (ICC = .301) with one another.

Quantity of Feedback: There was no significant difference in the average number of inline feedback comments between the two conditions (Fig. 5C; EvaluAId: $M = 2.167, SD = 2.725$; Baseline:

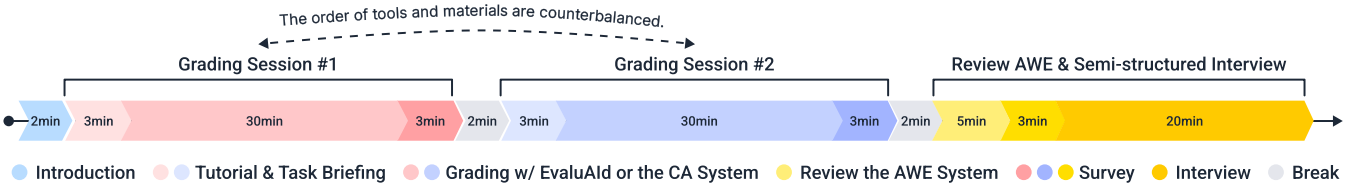


Figure 4: The within-subjects study procedure with TAs.

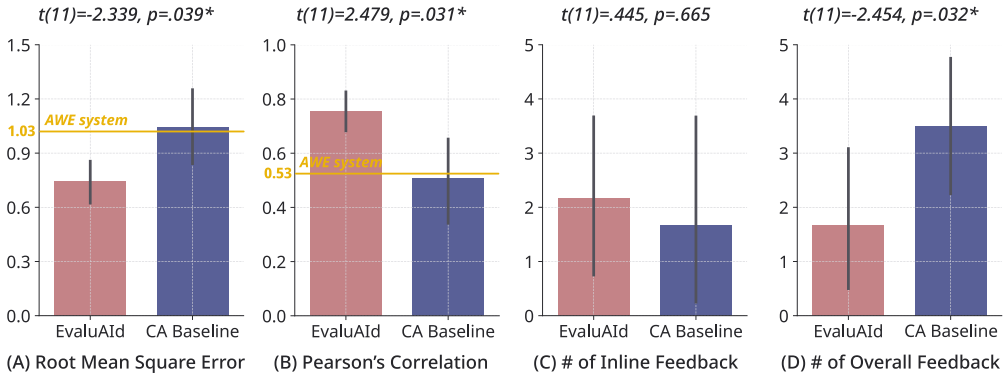


Figure 5: Bar plots comparing the performance of EvaluAId and the CA baseline across four metrics: (A) Root Mean Square Error (RMSE) relative to expert scores, (B) Pearson's correlation with expert scores, (C) number of inline feedback comments, and (D) number of overall feedback comments. Results show that EvaluAId achieved significantly lower differences (A) and higher correlation (B) with expert scores, while the CA baseline produced more overall feedback comments (D). The yellow line in (A) and (B) marks the performance of an LLM-based end-to-end AWE system.

$M = 1.667, SD = 3.447$; $t(11) = .445, p = .665$). However, participants using the baseline ($M = 3.500, SD = 2.431$) wrote significantly more overall feedback comments ($t(11) = -2.454, p = .032^{*}$) than those using EvaluAId ($M = 1.667, SD = 2.605$). This difference may stem from the CA baseline allowing participants to generate overall feedback more easily by simply inputting the student essays. In contrast, using EvaluAId required participants to spend additional time selecting relevant student content and rubric descriptors to synthesize their comments.

Quality of Feedback: Among independent rater evaluations of written comments, a significant difference was observed only in Clear Directions for Improvement, where feedback from EvaluAId ($M = 4.042, SD = 1.285$) was rated significantly higher ($t(23) = 1.798, p = .043^{*}$) than that from the baseline ($M = 3.292, SD = 1.546$). Interestingly, the Coefficient of Variation (CV) values for EvaluAId were consistently lower across all four quality metrics (e.g., Accuracy: 0.205 vs. 0.262, Criteria-Based: 0.245 vs. 0.296, Clear Directions for Improvement: 0.318 vs. 0.470, Supportive Tone: 0.298 vs. 0.339), suggesting that EvaluAId may help produce feedback of more consistent quality across participants.

5.2.2 RQ2.2: TA Perceptions of System Support. To answer RQ2, we conducted quantitative analysis using a repeated measures ANOVA to compare EvaluAId with the CA baseline and AWE across all metrics. Post hoc pairwise comparisons followed when significant effects were found. The results are summarized in Table 4.

Confidence and Satisfaction: Participants felt most confident in delivering assessment results to students when using EvaluAId ($M = 6.000, SD = 1.128$; $F = 12.931, p < .001^{***}$), especially compared to AWE ($M = 3.333, SD = 1.614$; post-hoc: $p < .001^{***}$). Also, TAs reported highest satisfaction with their final grading results when using EvaluAId ($M = 6.000, SD = .853$; $F = 10.577, p < .001^{***}$; post-hoc: $p = .004^{**}$) compared to the CA baseline ($M = 5.167, SD = 1.403$) and AWE ($M = 4.333, SD = 1.497$; post-hoc: $p = .004^{**}$).

Think-through and Collaboration: Participants perceived EvaluAId as significantly more effective in helping think through how to complete the task ($M = 6.500, SD = .674$; $F = 14.586, p < .001^{***}$) than the both systems (the CA baseline $M = 4.583, SD = 1.730$; post-hoc: $p = .002^{**}$; and AWE: $M = 3.667, SD = 1.923$; post-hoc: $p < .001^{***}$). TAs also experienced a strongest sense of collaboration when working with EvaluAId ($M = 5.500, SD = 1.382$; $F = 14.051, p < .001^{***}$), particularly compared to AWE ($M = 3.000, SD = 1.651$; post-hoc: $p < .001^{***}$).

Transparency and Control: EvaluAId provided greatest reported transparency in its decision-making process ($M = 5.750, SD = 1.712$; $F = 6.666, p = .005^{**}$) compared to CA ($M = 4.250, SD = 2.340$; post-hoc: $p = .069$) and AWE ($M = 3.333, SD = 1.875$; post-hoc: $p = .020^{*}$). TAs' perceived controllability over the AI was also highest with EvaluAId ($M = 5.083, SD = 1.881$; $F = 6.794, p = .005^{**}$), especially compared to AWE ($M = 3.000, SD = 1.595$; post-hoc: $p = .006^{**}$).

Table 4: Survey results under EvaluAId (*Ours*), the CA baseline, and AWE. We report means (M) and standard deviations (SD). Inferential statistics are from a one-way repeated-measures ANOVA across conditions, followed by Holm-corrected post-hoc paired t-tests (†: p<.10, *: p<.05, **: p<.01, ***: p<.001). All survey questions are presented in Supplementary Materials.

Metrics	EvaluAId (*Ours*)		CA		AWE		ANOVA		*Ours* vs. CA		*Ours* vs. AWE	
	M	*SD*	*M*	*SD*	*M*	*SD*	*F*	*p*	*t*	*p*	*t*	*p*
Confidence	6.000	1.128	5.250	1.658	3.333	1.614	12.931	.001***	1.621	.133	5.204	.001***
Satisfaction	6.000	0.853	5.167	1.403	4.333	1.497	10.577	.001***	4.022	.004**	4.212	.004**
Think-through	6.500	0.674	4.583	1.730	3.667	1.923	14.586	.001***	4.412	.002**	5.304	.001***
Collaboration	5.500	1.382	4.667	1.923	3.000	1.651	14.051	.001***	1.890	.085†	5.745	.001***
Transparency	5.750	1.712	4.350	2.340	3.333	1.875	6.666	.005**	2.413	.069†	3.345	.020*
Control	5.083	1.881	4.833	2.250	3.000	1.595	6.794	.005**	.333	.745	4.051	.006**
Trust	5.167	1.115	4.917	2.151	3.417	1.564	3.705	.041*	.370	.718	2.895	.044*
Reliability	5.333	1.231	4.833	2.250	3.417	1.676	4.146	.030*	.971	.352	3.027	.035*
Fairness	5.167	1.403	5.000	1.954	3.500	1.446	6.585	.006**	.352	.732	4.022	.006**
Justifiability	5.500	1.446	4.500	2.023	3.750	1.765	3.667	.042*	2.345	.078†	2.782	.054†

Trust and Reliability: Participants perceived EvaluAId as most trustworthy ($M = 5.167, SD = 1.115$; $F = 3.705, p = .041^{*}$), particularly better than AWE ($M = 3.417, SD = 1.564$; post-hoc: $p = .044^{*}$). Similar patterns emerged for reliability perceptions, with EvaluAId rated highest ($M = 5.333, SD = 1.231$; $F = 4.146, p = .030^{*}$) and significantly outperforming AWE ($M = 3.417, SD = 1.676$; post-hoc: $p = .035^{*}$).

Fairness and Justifiability: TAs perceived EvaluAId as leading to most fair outcomes for students ($M = 5.167, SD = 1.403$; $F = 6.586, p = .006^{**}$), significantly better than AWE ($M = 3.500, SD = 1.446$; post-hoc: $p = .006^{**}$), though not significantly different from the CA baseline ($M = 5.000, SD = 1.954$; post-hoc: $p = .732$). TAs also reported feeling most capable of justifying their grading decisions when disputes arose with EvaluAId ($M = 5.500, SD = 1.446$; $F = 3.667, p = .042^{*}$), although post-hoc tests did not reveal significant differences.

In summary, this study found that participants using EvaluAId achieved higher grading accuracy and consistency, produced more actionable feedback for student improvement, and reported greater satisfaction with their performance. They also noted that EvaluAId helped them think more carefully throughout the grading tasks. In the next section, we qualitatively examine EvaluAId from the perspectives of multiple grading stakeholders (i.e., namely instructors, TAs, and students) with the goal of understanding their perceived benefits of EvaluAId, uncovering potential mechanisms, and highlighting key considerations.

6 Study 3: Exploring Multi-Stakeholder Perspectives on Human-AI Collaborative Evaluation

Grading in university courses often involves multiple stakeholders: (1) TAs, who are typically responsible for grading; (2) instructors, who oversee the grading process and ensure consistency and fairness; and (3) students who are the recipients of feedback and scores and whose grade is directly impacted by the assessment. In the within-subjects study, we conducted semi-structured interviews with 12 TAs. Here, we further interviewed 6 university instructors and 6 university students to explore the perspectives of multiple stakeholders on EvaluAId's human-AI collaborative approach. We aim to investigate:

RQ3.1 What are the perceived benefits and challenges of EvaluAId's human-AI collaborative approach from the perspectives of TAs, instructors, and students?

RQ3.2 How might EvaluAId be integrated into classroom practice, what potential benefits does it offer for handling regrading requests, and what conditions need to be addressed?

6.1 Methodology

6.1.1 Participants. We recruited 6 instructors from universities across the United States through social media and Prolific (Table 5; I1-I6, 3 female, 3 male; ages ranging from 32 to 56, $M = 44.17$, $SD = 8.01$). They had extensive teaching experiences ranging from 6 to more than 10 years. Four were full-time professors, one was a full-time lecturer, and one was adjunct faculty. They come from diverse disciplines, including Art History, Computing, Comedy, Philosophy, Psychology, and Sociology. They all had experience designing and grading open-ended writing assignments such as essays, research papers, literature reviews, and reflective writings. Each instructor was compensated with $30.

We also recruited 6 university students (Table 6, S1-S6, 2 female and 4 male; ages ranging from 19 to 36, $M = 23.50$, $SD = 6.25$), including both undergraduate and graduate students, via social media and word-of-mouth. They came from a range of majors, such as Computer Science, Biology, Architecture, and Physics, and all had prior experience with writing assignments such as essays, discussion posts, and reflective writings. Each student was compensated with a $15 gift card.

6.1.2 Study Procedure. We conducted semi-structured interviews with 12 TAs in Study 2 (described in Section 5.1.4). For instructor and student participants, we conducted one-on-one interviews via Zoom, each lasting about an hour. At the start of each session, we introduced the study's background and goals. Participants then provided verbal consent for audio and video recording and completed a

Table 5: Demographic information of university instructors in Study 3.

ID	Gender	Age	Teaching Experience	Field	Class Size	AI Grading
I1	Female	48	> 10 years	Art History	< 20	Not yet but interested
I2	Female	56	> 10 years	Computing	20-49	No
I3	Male	32	6-10 years	Comedy	20-49	No
I4	Female	46	6-10 years	Philosophy	20-49	Used to detect AI in writing
I5	Male	41	> 10 years	Psychology	< 20	Not yet but interested
I6	Male	42	> 10 years	Sociology	< 20	No

Table 6: Demographic information of university students in Study 3.

ID	Gender	Age	Grade	Major Field
S1	Male	23	Senior	Computer Science
S2	Male	21	Senior	Computer Science
S3	Female	21	Senior	Molecular, Cellular, and Developmental Biology
S4	Male	19	Sophomore	Physics
S5	Male	36	Graduate	Biomedical Sciences
S6	Female	21	Senior	Architecture

demographic survey. We engaged participants in a warm-up activity where they shared their experiences with writing assignments and initial impressions of the AI in grading.

Next, we presented three systems in order (the CA system, the AWE system, and EvaluAId) by walking through key features and showing anonymized screen recording clips of TAs using each system from Study 2. Each clip combined representative segments of TAs' grading behavior from Study 2 and showcased all system features and workflows. After each presentation, participants discussed the potential benefits and concerns of using the system in grading. Following all three demonstrations, we held a 30-minute semi-structured conversation inviting participants to compare the systems and reflect on their differences.

For instructors, we focused on identifying their favorite features across systems, gauging their attitudes toward EvaluAId's human-AI collaborative grading approach compared to CA and AWE, and envisioning how EvaluAId could be integrated into their classrooms. For students, we explored their acceptance of and attitudes towards EvaluAId's collaborative approach, its potential impact on their learning, and their perceived benefits and concerns across the three systems. The full interview protocol is provided in Supplementary Materials.

6.2 Analysis and Findings

To analyze interview transcripts, we followed established open-coding protocols [8, 68]. Two researchers independently coded the transcripts, then discussed, reached a consensus, and created a consolidated codebook. This codebook was then used for thematic analysis to identify emerging topics from the interviews. The entire research team collectively reviewed the coding outcomes to refine high-level themes.

6.2.1 Evidence Identification: Benefits and Challenges. Here, we report instructors', TAs', and students' perceived benefits and challenges regarding EvaluAId' approach to evidence identification.

KF1: Criteria decomposition and sentence highlighting enable evidence-driven judgment: Both TAs and instructors agreed that consideration-based, sentence-level highlighting can help graders effectively locate rubric-aligned evidence (T1-10, I1, I3) by breaking apart each criteria into considerations and filtering out irrelevant sentences (T3, T5). The benefit was especially pronounced for poorly structured essays (T1, T7). Compared to chatbot responses, participants viewed in-place highlights as more trustworthy because they preserved context (T4, T5, I3, I4, S6). As I3 noted, "*It shows you the context around the thing... It's just highlighting instead of rewriting.*" Participants also cautioned that for subjective criteria with diffuse evidence (e.g., comparisons spread across paragraphs), highlights should be treated as prompts rather than proofs; otherwise graders might miss cross-paragraph reasoning (T4).

Our findings revealed differing perspectives on the transparency of showing highlighted sentences. Instructors preferred not to share highlights, expressing worrying that it might "*give students more feedback than they actually want*" (I1). By contrast, students regarded scores paired with highlights as "*very specific feedback*" (S2, S3) that helped pinpoint mistakes. This divergence motivates an important direction for future work: what is the optimal level of detail to provide during the scoring and feedback process.

KF2: Turning rubric into guiding questions promotes reflection and builds confidence: In EvaluAId, each criterion is decomposed into guiding questions (considerations) that prompt TAs to evaluate the criterion along multiple facets. TA users reported that these considerations encouraged more thoughtful grading (T3, T5-7, T10), consistent with our observation that EvaluAId supports deeper think-through in the within-subjects study (§5.2.2). For example, T3 and T5 filtered sentences consideration-by-consideration, moving from surface-level checks (e.g., number of concepts defined) to more nuanced aspects (e.g., quality of explanation). This process helped them gather evidence gradually to confirm their grading decisions: "*[Each consideration] will give me some highlighted evidence*

Table 7: Summary of insights from TAs, instructors, and students on EvaluAID's human-AI collaborative approach.

Key Dimensions	TAs	Instructors	Students
Evidence Identification: **B**enefits and **C**hallenges	**B1**. TAs appreciate that sentence highlighting can help them easily locate key points and allocate attention. **(KF1)** **B2**. TAs prefer in-place highlights over AI summaries, because they are more trustable. **(KF1)** **B3**. TAs perceive being more thoughtful and guided by considerations in EvaluAID. **(KF2)** **C1**. TAs have concerns about the AI's reliability in detecting sentences for very subjective criteria. **(KF1)**	**B1**. Instructors agree that highlighting can help graders locate key points easier. **(KF1)** **B2**. Instructors trust in-place highlights more than AI summaries or AI quotes, because in-place highlights keep the original context. **(KF1)** **C1**. Instructors have concerns about LLM missing important sentences for making grading decisions. **(KF1)** **C2**. Instructors are reluctant to deliver highlights as part of feedback to students. **(KF1)**	**B1**. Students appreciate the deeper engagement of graders with their work encouraged by sentence highlighting and considerations. **(KF2)** **B2**. Students trust in-place highlights more than AI-extracted quotes, because they fear hallucinations in the AI results. **(KF1)** **C1**. In contrast to instructors' attitudes, students prefer receiving highlighted sentences as part of the feedback. **(KF1)**
Comparative Judgment: **B**enefits and **C**hallenges	**B1**. TAs appreciate that benchmarking visualization can help them quickly spot similar assignments and make pairwise comparison. **(KF3)** **B2**. TA appreciate that the AI analysis of strengths and weaknesses can help them interpret the quality of a submission and reflect on their judgment. **(KF3)** **C1**. TAs have concerns about the reliability of the similarity calculation when students' writing topics are very diverse. **(KF4)**	**B1**. Instructors propose iterative benchmarking facilitated by EvaluAID: they tend to treat each essay as a benchmark to ensure fair attention. **(KF4)** **B2**. Instructors believe that EvaluAID can elicit deeper reflection on grading standards through the benchmarking and holistic view of writings in the grading canvas. **(KF3)** **C1**. Instructors have concerns about unfair comparisons with outlier submissions. **(KF4)**	**C1**. Students have concerns that graders might select inappropriate essay (such as outliers) as the benchmark. **(KF4)** **C2**. Students care about whether graders maintain consistent standards during comparative grading. **(KF4)**
Feedback Composition: **B**enefits and **C**hallenges	**B1**. TAs are satisfied with the quality of the AI feedback draft in EvaluAID, as they are concise, accurate, and concrete. **(KF5)** **B2**. TAs perceive that the feedback composition mechanism can guide them to reflect on their judgment and write thoughtful feedback with justifications. **(KF5)**	**B1**. Instructors believe that the feedback composition mechanism in EvaluAID can make graders intentionally incorporate students' work into the feedback. **(KF5)** **B2**. Instructors expect that the personalized feedback supported by EvaluAID will strengthen their connections with students. **(KF5)** **B3**. Instructors believe that the similarity-to-AI score can help graders manage their reliance on the AI. **(KF6)** **C1**. Instructors see a need to balance structured feedback writing with flexibility to request AI help. **(KF5)**	**B1**. Students perceived that the AI feedback draft in EvaluAID are informative for revision. **(KF5)** **C1**. Students appreciate the human effort in selecting evidences but expect more human input to build on the AI draft. **(KF6)**
Integration into Classroom: **B**enefits and **C**hallenges	**B1**. TAs believe that grading based on considerations and writing feedback by quoting students' work and rubric descriptions can help them articulate grading rationale when regrading is requested. **(KF7)** **B2**. By breaking grading into steps, EvaluAID can give TAs greater control and help them better understand how the system works. **(KF8)**	**B1**. Instructors believe that EvaluAID can help resolve regrade requests because it can guide graders to form explicit rationale and maintain accountability. **(KF7)** **B2**. Instructors see EvaluAID as a promising tool for training TAs. **(KF8)** **C1**. Instructors want guidance on configuring or customizing the AI support. **(KF8)** **C2**. EvaluAID should fit in, not replace, instructor-student interactions. **(KF8)**	**B1**. Students note that sentence highlighting can reduce their regrade requests caused by human oversight and make graders' rationale more transparent. **(KF7)** **B2**. Students note that having graders specify rubric snippets when writing feedback can reduce grade disputes by leaving less room to argue for points. **(KF7)**

... I will first have a feeling ... then I need some evidence ... if I click the consideration, I could reassure myself." (T5)

Students also valued this practice, perceiving that graders were actively engaging with their writing: "*...the instructor is interacting more deeply with [my work], because they're reading and associating writing snippets with [considerations] and making their own judgment.*" (S4) In contrast, after observing CA use, S4 felt the grader was "*just scanning... to form a broad impression, then filling in details without judging.*" For students, agency and authenticity were just as valuable as the pedagogical content of the feedback.

6.2.2 Comparative Judgment: Benefits and Challenges. Here, we report instructors', TAs', and students' perceived benefits and challenges of EvaluAID' human-AI collaborative approach regarding comparative judgment.

KF3: Adaptive benchmarking and grading Canvas support efficient, deliberated comparative judgment: TAs indicated that the similarity color-coding in EvaluAId helped them quickly identify pairs for comparison (T1, T2, T4, T6, T8). As T6 explained, "*When I compare two essays, it's easier to see who is doing better and [identify] the problems.*" TAs also used the AI analysis of an essay's strengths and weaknesses compared to the chosen baseline to actively interpret writing quality (T1). While EvaluAId improves grading efficiency, instructors noted that it also prompted deep reflection on their grading standards (or consistency) through its color-coding, comparative analysis, and holistic view of all writings on the grading canvas. As I3 described: "*I would look at this dark blue essay that got 2 points and the one that got 4 points, and think: why do I rate this one a 2 and that one a 4, even though the 3-point essay is very similar to both? ... This keeps me in check when I'm doing a marathon of grading: what outside factors might be affecting my mood and whether I was being fair.*" This kind of reflection on consistency is important because it helps ensure that grading remains fair and aligned with shared standards, even when evaluators face fatigue or subjective bias.

KF4: Iterative benchmarking can potentially support fairness in comparative judgment: Participants raised concerns about fairness in peer comparative judgment. For instance, students worried that graders might select inappropriate benchmarks or apply inconsistent standards (S4). Instructors similarly noted the risk of being influenced by outlier assignments (e.g., unusually strong essays), which could lead to unfair comparisons (I5). They advocated treating every essay as a potential baseline to distribute attention equitably (I3). Participants also expressed concerns over the reliability of content similarity when student essays vary widely in topics (T3). EvaluAId's flexible baseline selection and holistic views of all essays in the canvas can encourage such iterative benchmarking workflow such as revisiting the baseline, making pairwise comparisons, and checking for inconsistency in standards. This points to a broader question around the right balance of peer comparison versus independent grading. Though interventions such as those in EvaluAId can help to reduce the effort of cross-referencing, there remains a risk that metrics like similarity can cause priming or lead instructors to unwittingly fixate on one example.

6.2.3 Feedback Composition: Benefits and Challenges. In Study 2 (§5.2.1), we found that EvaluAId helped TAs produce more actionable feedback. Here, we provide further evidence, reveal the potential mechanisms, and highlight considerations raised by instructors, TAs, and students.

KF5: Evidence-first workflows support personalized, informative feedback composition: TAs and students were satisfied with the quality of AI-generated feedback drafts, describing them as "*concise, accurate, and concrete*" (T1, T5, T7, S2), aligned with our findings in §4.5. Students appreciated that EvaluAId guides graders to tie feedback to rubric considerations and essay snippets: "*I really like how this system allows instructors to give suggestions based on [essay] snippets or guiding questions... It's going to be more informative for students to address certain points in future writing.*" (S4) TAs also noted that the structured feedback-writing steps reminded them to justify their grading decisions (T9).

Instructors similarly observed that the scaffolding process can encourage graders to more consciously write personalized feedback: "*I would pull out essay snippets that I really want to give constructive feedback on, knowing that at the end I'd be generating feedback and could plug those snippets in and edit them ... Whereas now I'm not that intentional. ... This [EvaluAId] would make me more intentional about identifying which parts of the essay to incorporate into feedback.*" (I5) They valued how EvaluAId's scaffolding and the AI suggestions prompted reflection on feedback quality when incorporating the AI content (I3), and believed that more personalized feedback can help sustain meaningful instructor-student connections even when the AI assists grading (I3).

At the same time, instructors and TAs emphasized balancing structured guidance with flexibility in feedback composition. Compared to EvaluAId's structured approach, instructors found that the chatbot in the CA system sometimes allowed more creative exploration, because "*it's much less on rails and can generate anything.*" (I5) This suggests the potential benefits of combining conversational interactions with structured, snippet-grounded scaffolding for feedback writing.

KF6: Human judgment and visible effort matters in feedback composition: TAs described EvaluAId as better supporting their responsibility to make independent judgments than CA and AWE (T1-4, T7, T12). T4 perceived greater agency in EvaluAId because "*it's just providing tools for graders to present their own judgment instead of giving an answer.*" Students agreed with this view: "*In [CA], they [graders] are being influenced by the AI because they can chat with it. Whereas in EvaluAId, they're making the decision first without any AI.*" (S2) Many participants regarded AWE as affording the least agency, describing it as "*making the human grader a side person in the grading*" (T7, T9, I3-5) and "*giving graders the temptation to be lazy.*" (I1, I3, I4, S2, S4)

Instructors emphasized contributing original comments to preserve authenticity and a sense of connection with student work: "*I would make it a point to provide original feedback for each submission, just so that I am still having that...intimate interaction with the essay.*" (I3) Students similarly valued visible human effort in feedback, preferring human-edited AI drafts for tone and correctness (S2). The similarity-to-AI indicator in EvaluAId was viewed as a useful guardrail: "*I really love that similarity-to-AI feedback score to let me know how heavily am I leaning on this generated content to provide feedback?*" (I5) The positive responses to EvaluAId are encouraging, but there remains a central tension in the AI-supported evaluation between scalability of grading approaches and close connections to students.

6.2.4 Integration into Classroom. Lastly, we report participants' perspectives on integrating EvaluAId into classrooms.

KF7: Rubric- and text-linked grading rationales help resolve regrade disputes: When discussing classroom integration, participants emphasized how EvaluAId can help address grading conflicts. TAs reported that the explicit rationale and linked evidence encouraged in EvaluAId can make disputes easier to resolve. For example, considerations can help with justification: "*For grading, the most important part for me is to find some justification for my grading. So this...gave me several considerations I can have for the student.*" (T5)

Linking feedback to rubrics and student content further helps them justify their decisions (T9). Instructors likewise noted the value in regrade situations: rather than relying on memory weeks later, they appreciated having a record of evidence linked to criteria: "*If a student has a dispute...I could click [a criteria] and be like here is [the evidence] specifically...it's not me having to recall what happened when I interacted.*" (I3)

Students echoed these views, reporting greater trust when feedback contained explicit rubric references and supporting evidence. Such detail reduced both the desire and the grounds to dispute grades: "*If the instructor is giving suggestions of certain parts of the essay associated with certain rubric, I feel it more difficult to find any 'sweet spots' to argue...because it already lists everything out for you, so I would trust their judgment a bit more.*" (S4) Students also noted that the AI highlights can help graders recognize their efforts more fully, reduce disputes stemming from overlooked points (S4), and make graders' rationale more transparent when questions arise.

KF8: Integrating EvaluAId into classrooms requires transparency, adaptation, and training: Transparency was a recurring theme. Instructors stressed being explicit about AI use to avoid confusion or mistrust from students and parents: "*We'll have a lot of parental concerns about AI grading my student. ... There will be multiple stages of experiencing what it actually is versus the rumor of what it is.*" (I1) Compared to the black-box AWE, instructors and TAs felt EvaluAId' processes are easier to explain. Instructors also emphasized adaptation. They wanted guidance on configuring or customizing the AI support to match their grading styles and pedagogical goals (I3) and reiterated that the system should fit in, not replace, instructor-student interactions (I1): "*I spend a lot of time doing one on one evaluation with those students. and I think this would just be one more tool for that. But I do think you would have to have conversations with students about the system, what the system is finding and how you can use what the system is telling you to improve.*"

Based our observations, onboarding is needed to help TAs familiar with EvaluAId's interactions, such as the drag-and-drop feature (T3, T5). Some TAs also noted that the benchmarking color coding across similar essays could be misinterpreted as implying similar quality (T4, T8). Since human judgment is still required to verify such comparisons, training should guide TAs on how to appropriately interpret benchmarking results and avoid misuse, such as equating writing similarity with grading equivalence. Finally, instructors also saw EvaluAId as a valuable TA training tool. By surfacing criteria-based highlights and enabling benchmarking, it can help novice TAs apply rubrics consistently and recognize high-quality work. As I1 noted: "*If I was putting together a system to help teach people how to grade, this would be a really great one...especially if you had a benchmark paper that you could insert into it as the ideal.*"

7 Discussion

In Study 3, we highlighted the important role of thoughtfulness facilitated by EvaluAId in the grading process and identified key considerations raised by instructors, TAs, and students. Here, we discuss how to further foster thoughtfulness in human-AI collaborative grading, what constitutes effective human-AI collaboration in grading, and considerations for scaling EvaluAId in real-world classroom settings.

7.1 Fostering Thoughtfulness in Human-AI Collaborative Grading

We use *thoughtfulness* to describe a metacognitive mode of grading characterized by deliberate, evidence-first reasoning and reflective judgment. Drawing from cognitive and educational psychology, this notion aligns with three well-established constructs. First, metacognitive monitoring [19] highlights graders' awareness of their own understanding and uncertainty when interpreting student work so they can recognize when more evidence or clarification is needed. Second, reflective judgment [42] emphasizes reasoning about claims whose correctness cannot be verified absolutely, but must be justified through evidence and criteria. Third, deliberate reasoning under uncertainty [40] contrasts slow, analytic thought with fast, heuristic judgment, underscoring the importance of pausing to weigh alternative interpretations before deciding. In rubric-based grading, such thoughtfulness manifests as actively interpreting textual evidence, evaluating its fit with rubric levels, and articulating justifications that connect evaluation with improvement.

EvaluAId fosters thoughtfulness by reducing low-level search effort and reallocating attention toward these higher-order cognitive processes. Instead of spending time scrolling and locating relevant content, graders can focus on determining how candidate snippets align with rubric descriptors, calibrating against comparable work, and composing feedback that explains both why a score is warranted and how revision could improve performance. End-to-end automation, however, risks reducing such thoughtfulness because automation can induce complacency [62] and "out-of-the-loop" [15] effects that reduce active monitoring and critical judgment. In educational autograding contexts [35], such automated outputs anchor behavior and invite superficial strategies (such as counting words [63]) rather than deliberation.

The mechanisms that foster thoughtfulness in EvaluAId reflect the four design goals introduced earlier in Table 2: interactive evidence mapping enables deliberate evidence interpretation **[DG1]**; adaptive benchmarking encourages reflective calibration through comparison **[DG2]**; scaffolded feedback synthesis supports articulated justification and improvement guidance **[DG3]**; and human-in-the-loop control preserves agency and accountability during evaluation **[DG4]**. Together, these goals constitute theoretically informed design principles that extend beyond our specific system to broader contexts of human-AI collaboration in evaluative tasks. We call for future research to examine how these principles generalize across domains (e.g., peer review, hiring, or creative assessment) and how thoughtful human-AI interaction can be operationalized, measured, and sustained in real-world settings.

7.2 What Constitutes Effective Human-AI Collaboration in Grading

In this paper, we reposition the AI not as an autonomous grader but as an on-demand collaborator that provides specific, targeted support to human graders. Our studies offer initial answers, but not full accounts, to broader questions about the effectiveness of human-AI collaboration in grading. In this section, we draw on Holstein et al.'s framework for human-AI hybrid adaptivity [34], which describes how humans and AI can augment one another through perceptual, action, and decision augmentation, to discuss

what constitutes "effective collaboration" in our setting. By the end, we surface open questions that our studies only begin to address and encourage future work.

7.2.1 Perceptual Augmentation. EvaluAId primarily augments perception by helping graders notice rubric-relevant evidence and patterns across essays through guided highlighting, baseline comparison, and similarity views. TAs reported that these features "*helped them think through*" grading and rated EvaluAId as more transparent and justifiable than the other conditions. At the same time, all stakeholders emphasized the need to double check AI surfaced evidence for nuanced criteria, which suggests that effective perceptual augmentation makes evidence more visible without fixing the interpretation or removing the need for critical reading.

7.2.2 Action and Decision Augmentation. In EvaluAId, the AI extends graders' action space by structuring difficult grading moves into manageable steps. The system does not output final scores. Instead, it shapes TAs' decision making through intermediate artifacts such as highlighted evidence and baseline comparisons. On the other hand, TAs extend the AI's action space by choosing guiding questions, baselines, and excerpts that adapt the generic model to course-specific standards and local context. Our quantitative results indicate that this human-AI hybrid scaffolding improved grading accuracy and inter-rater consistency, led to more actionable feedback, and was favored by other stakeholders in Study 3. Prior work often bypasses this kind of action augmentation by directly delivering grading decisions to graders [29, 51, 57, 67, 81, 93]. Our work highlights the importance and effectiveness of action augmentation in human-AI collaborative grading and points to a form of decision augmentation where the AI helps graders articulate and apply their own decision policies, while responsibility and final authority remain with human graders.

7.2.3 Open Questions and Future Work. Several open questions remain: What are the cognitive and affective experiences of TAs when collaborating with AI? How do TAs develop trust and calibrate their reliance on AI suggestions over time? What individual differences moderate the effectiveness of collaboration? Our findings provide early evidence about cognitive and affective experiences (greater confidence, but also moments of doubt about over-reliance) and emerging trust calibration strategies (accepting AI help on local, checkable aspects and being more critical for subjective criteria). Future work should examine these processes over longer deployments (we discuss considerations for deploying EvaluAId in real-world settings in §7.3) and combine behavioral logs with cognitive, affective, and trust measures [38, 80, 82]. We also encourage future work to systematically model how grading experience, domain expertise, and AI literacy moderate the effectiveness of collaboration, and to examine under what conditions graders appropriately accept versus critically evaluate AI-generated highlights and feedback.

7.3 Scaling and Deploying EvaluAId in Real-World Classroom Settings

A key question for scaling EvaluAId is how to integrate it into everyday classrooms. Here we discuss four considerations derived from the design of EvaluAId and the findings from Study 3.

7.3.1 Consideration 1: Conflict-Resolution Workflows and Audit Trails. Our findings suggest EvaluAId can potentially help instructors, TAs, and students resolve grading conflicts: it prompts evidence-first rationales for TAs, gives instructors traceable justifications, and offers students evidence-based feedback that reduces regrade requests (KF7). Prior work also shows that detailed, evidence-based feedback increases students' acceptance of grading decisions [60]. In contrast, students frequently overestimate an autograder's chance of marking correct answers as wrong [35]. This distinction matters in classrooms, where disputes consume time, divert resources from teaching [26], and strain student-teacher relationships [46]. When deploying EvaluAId, we need to further introduce a regrade/dispute workflow: (i) a "regrade mode" that exports a shareable evidence packet (e.g., highlighted excerpts, linked rubric considerations) and (ii) versioned, role-based audit trails that record who selected evidence, edited AI drafts, and finalized scores.

7.3.2 Consideration 2: Policy and Parity to Promote Mutual Trust. Both instructors and students noted a sense of inequity if the teaching team were to use the AI for grading while restricting students from using the AI for writing. To ease this tension, we need to adopt an AI-use parity policy when deploying EvaluAId that specifies (i) what AI assistance is permitted for graders (e.g., evidence highlighting only, no auto-scoring), (ii) what AI assistance is permitted for students (e.g., language polish with disclosure), (iii) required disclosure statements for both parties (brief "AI involvement" notes), and (iv) non-AI fallback options for students who opt out. Clear policy and symmetric expectations can help promote mutual trust.

7.3.3 Consideration 3: Transparency and Course-Level Communication. Transparency is essential for trust between instructors and students and for addressing parental concerns (KF8). Students expect disclosure when the AI is used, and instructors must explain how it works in their course. To communicate transparently, instructors should walk students through example graded assignments to demonstrate how evidence and rationale are generated. When class size permits, instructors can also hold one-on-one conversations with students to discuss the feedback they receive from EvaluAId (KF8). Through these conversations, instructors can continually evaluate how EvaluAId is perceived by students and adjust guidelines for graders to address emerging concerns. A handbook of EvaluAId for students will also be useful: (i) a one-page "How your work is graded with EvaluAId" statement; (ii) per-assignment feedback reports that include highlighted evidence, mapped considerations, baseline comparisons, and human-authored justifications; and (iii) plain-language explanations of "similarity" metrics and their limits.

7.3.4 Consideration 4: Building AI Literacy for Graders and Students. Finally, successful deployment requires raising AI literacy among instructors, TAs, and students. Instructors need enough understanding of model behavior to adjust assignments, prompts, and materials; TAs need the know-how to steer models and spot errors or bias; and students need literacy to interpret AI-mediated feedback and use the AI judiciously. When deploying EvaluAId, we need to offer lightweight training and scaffolds: (i) short, role-specific onboarding (instructor/TA/student) with sandbox "shadow grading" before live use; (ii) built-in micro-tutorials and "why this suggestion?" tooltips

that reveal retrieval sources and limitations; (iii) periodic calibration sessions using exemplars to detect drift in standards; and (iv) student activities that critique AI-generated feedback, cultivating informed acceptance of feedback rather than blind reliance.

7.4 Limitations & Future Work

Our system design has several limitations. It presently targets three cognitive processes, namely, evidence identification, comparative judgment, and feedback composition. Grading, however, involves additional mechanisms and biases (e.g., representativeness, availability, anchoring [78, 79]) that can shape attention and decisions. Future work could add bias-aware features, such as blinded ordering, uncertainty cues, and reflective checkpoints. Second, EvaluAId currently operates on text (both essays and RAG sources). Many classroom submissions include figures, screenshots, diagrams, or tables; extending to multimodal inputs would require layout-aware parsing, image/figure-text alignment, and multimodal retrieval grounding. Third, collaboration features are limited. Future iterations could support graders with shared baselines, disagreement-resolution workflows, and instructor dashboards to monitor drift, coverage, and cohort progress. Fourth, while rubrics are a common and powerful grading tool, instructors in our study noted challenges in designing good rubrics particularly for novices. Future work could design tools for rubric design and refinement.

Our study methodology also has limitations. The within-subjects study focused on one open-ended assignment in an HCI course at a single institution with modest sample sizes. While we computed statistical significance, the tests may be underpowered; thus, we do not claim the results to be conclusive but should be interpreted as promising but preliminary. To further evaluate the effectiveness of our design, we plan to conduct a large-scale classroom deployment in collaboration with instructors and TAs at our own and partner universities. Although Study 3 included instructors from multiple fields, future work should further quantitatively examine EvaluAId across courses, disciplines, and rubric granularities with a larger sample size. Second, the qualitative and retrospective nature of our multi-stakeholder interviews limits generalizability. Students and instructors only viewed videos, which limits ecological validity compared to using EvaluAId and actually receiving grading results and feedback as a student. Third, our studies were short-term. Longitudinal deployments are needed to evaluate sustained calibration, grader workload and satisfaction, student learning outcomes, fairness, and handling of regrade requests. For example, future work could examine how graders justify scores with linked evidence during regrade disputes, and whether EvaluAId helps maintain consistent and accountable decisions over time.

8 Conclusion

This paper reframes the AI's role from autonomous grader to on-demand collaborator that can provide specific, targeted support. We introduced EvaluAId, which surfaces rubric-aligned evidence, supports adaptive benchmarking and self-calibration, and scaffolds human-steered, retrieval-grounded feedback, informed by a formative study with TAs. Across a within-subjects study with 12 TAs, EvaluAId improved alignment with expert scores and increased graders' satisfaction and reflective judgment relative to an interactive rubric+LLM chatbot and an LLM-based AWE. Semi-structured interviews with TAs, instructors, and students, underscored the value of thoughtfulness supported by EvaluAId while surfacing practical considerations for integration into classroom. Together, these findings advocate for deliberate, evidence-first, human-in-the-loop evaluation as an alternative to end-to-end automation.

Acknowledgments

We sincerely thank all TAs, instructors, and students who participated in our study for generously sharing their insights. We also thank the reviewers for their valuable comments and suggestions.

References

[1] Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering. *Transactions of the Association for Computational Linguistics* 12 (May 2024), 681–699.

[2] Omar Alsaiari, Nilufar Baghaei, Hatim Lahza, Jason Lodge, Marie Boden, and Hassan Khosravi. 2024. Emotionally Enriched Feedback via Generative AI. arXiv:2410.15077 [cs]

[3] Saleema Amershi, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, Eric Horvitz, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, and Paul N. Bennett. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM Press, Glasgow, Scotland Uk, 1–13. doi:10.1145/3290605.3300233

[4] Pengcheng An, Kenneth Holstein, Bernice d'Anjou, Berry Eggen, and Saskia Bakker. 2020. The TA Framework: Designing Real-Time Teaching Augmentation for K-12 Classrooms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–17. doi:10.1145/3313831.3376277

[5] Yigal Attali and Jill Burstein. 2006. Automated Essay Scoring With E-Rater® V.2. *The Journal of Technology, Learning and Assessment* 4, 3 (Feb. 2006), 1–31.

[6] Giovanna Badia. 2019. Holistic or Analytic Rubrics? Grading Information Literacy Instruction. *College & Undergraduate Libraries* 26, 2 (April 2019), 109–116. doi:10.1080/10691316.2019.1638081

[7] Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-Defined AI Personas for On-Demand Feedback Generation. arXiv:2309.10433 [cs] doi:10.1145/3613904.3642406

[8] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. doi:10.1191/1478088706qp063oa

[9] Julia Cambre, Scott Klemmer, and Chinmay Kulkarni. 2018. Juxtapeer: Comparative Peer Review Yields Higher Quality Feedback and Promotes Deeper Reflection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–13. doi:10.1145/3173574.3173868

[10] Ten Cate and Olle Th J. 2013. Why Receiving Feedback Collides with Self Determination. *Adv in Health Sci Educ* 18, 4 (Oct. 2013), 845–849. doi:10.1007/s10459-012-9401-0

[11] Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 15607–15631. doi:10.18653/v1/2023.acl-long.870

[12] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376638

[13] Paul Deane. 2013. On the Relation between Automated Essay Scoring and Modern Views of the Writing Construct. *Assessing Writing* 18, 1 (Jan. 2013), 7–24. doi:10.1016/j.asw.2012.10.002

[14] Ravit Dotan, Lisa S. Parker, and John Radzilowicz. 2024. Responsible Adoption of Generative AI in Higher Education: Developing a "Points to Consider" Approach Based on Faculty Perspectives. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 2033–2046. doi:10.1145/3630106.3659023

[15] Mica R. Endsley and Esin O. Kiris. 1995. The Out-of-the-Loop Performance Problem and Level of Control in Automation. *Hum Factors* 37, 2 (June 1995), 381–394. doi:10.1518/001872095779064555

[16] 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 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3654777.3676390

[17] Antonio Ferrara, Francesco Bonchi, Francesco Fabbri, Fariba Karimi, and Claudia Wagner. 2024. Bias-Aware Ranking from Pairwise Comparisons. *Data Min Knowl Disc* 38, 4 (July 2024), 2062–2086. doi:10.1007/s10618-024-01024-z

[18] Brianna Finnegan. 2024. *Teacher Use Of Rubrics To Assess Claim, Evidence And Reasonings In The High School Science Classroom.* Ph. D. Dissertation. University of Northern Iowa.

[19] John H. Flavell. 1979. Metacognition and Cognitive Monitoring: A New Area of Cognitive–Developmental Inquiry. *American psychologist* 34, 10 (1979), 906.

[20] Atta Gebril and Lia Plakans. 2014. Assembling Validity Evidence for Assessing Academic Writing: Rater Reactions to Integrated Tasks. *Assessing Writing* 21 (July 2014), 56–73. doi:10.1016/j.asw.2014.03.002

[21] Sharan A. Gibson and Pamela Ross. 2016. Teachers' Professional Noticing. *Theory Into Practice* 55, 3 (July 2016), 180–188. doi:10.1080/00405841.2016.1173996

[22] Grammarly. 2025. Grammarly: Free AI Writing Assistance. https://www.grammarly.com/.

[23] Shirley Gregor, Leona Chandra Kruse, and Stefan Seidel. 2020. Research Perspectives: The Anatomy of a Design Principle. *Journal of the Association for Information Systems* 21, 6 (Nov. 2020), 1–49. doi:10.17705/1jais.00649

[24] Douglas Grimes and Mark Warschauer. 2010. Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation. *The Journal of Technology, Learning and Assessment* 8, 6 (2010), 1–44.

[25] Qingyu Guo, Chao Zhang, Hanfang Lyu, Zhenhui Peng, and Xiaojuan Ma. 2023. What Makes Creators Engage with Online Critiques? Understanding the Role of Artifacts' Creation Stage, Characteristics of Community Comments, and Their Interactions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3544548.3581054

[26] Vidar Gynnild. 2011. Student appeals of grades: a comparative study of university policies and practices. *Assessment in Education: Principles, Policy & Practice* 18, 1 (2011), 41–57.

[27] Erin Hall, Mohammed Seyam, and Daniel Dunlap. 2024. Exploring Explainability and Transparency in Automated Essay Scoring Systems: A User-Centered Evaluation. In *Learning and Collaboration Technologies: 11th International Conference, LCT 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Washington, DC, USA, June 29–July 4, 2024, Proceedings, Part III*. Springer-Verlag, Berlin, Heidelberg, 266–282. doi:10.1007/978-3-031-61691-4_18

[28] Bingyi Han, Simon Coghlan, George Buchanan, and Dana McKay. 2024. Who Is Helping Whom? Student Concerns about AI- Teacher Collaboration in Higher Education Classrooms. arXiv:2412.14469 [cs] doi:10.48550/arXiv.2412.14469

[29] Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. 2023. FABRIC: Automated Scoring and Feedback Generation for Essays. arXiv:2310.05191 [cs] doi:10.48550/arXiv.2310.05191

[30] Maralee Harrell. 2005. Grading According to a Rubric. *Teaching Philosophy* 28, 1 (2005), 3–15. doi:10.5840/teachphil200528111

[31] Emma Harvey, Allison Koenecke, and Rene F. Kizilcec. 2025. "Don't Forget the Teachers": Towards an Educator-Centered Understanding of Harms from Large Language Models in Education. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3706598.3713210

[32] Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 13806–13834. doi:10.18653/v1/2024.acl-long.745

[33] John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research* 77, 1 (March 2007), 81–112. doi:10.3102/003465430298487

[34] Kenneth Holstein, Vincent Aleven, and Nikol Rummel. 2020. A Conceptual Framework for Human–AI Hybrid Adaptivity in Education. *Artificial Intelligence in Education* 12163 (June 2020), 240–254. doi:10.1007/978-3-030-52237-7_20

[35] Silas Hsu, Tiffany Wenting Li, Zhilin Zhang, Max Fowler, Craig Zilles, and Karrie Karahalios. 2021. Attitudes Surrounding an Imperfect AI Autograder. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3411764.3445424

[36] Jussi S. Jauhiainen and Agustín Garagorry Guerra. 2024. Evaluating Students' Open-Ended Written Responses with LLMs: Using the RAG Framework for GPT-3.5, GPT-4, Claude-3, and Mistral-Large. arXiv:2405.05444 [cs] doi:10.48550/arXiv.2405.05444

[37] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12 (March 2023), 248:1–248:38. doi:10.1145/3571730

[38] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (March 2000), 53–71. doi:10.1207/S15327566IJCE0401_04

[39] Anders Jonsson and Gunilla Svingby. 2007. The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences. *Educational Research Review* 2, 2 (Jan. 2007), 130–144. doi:10.1016/j.edurev.2007.05.002

[40] Daniel Kahneman. 2011. *Thinking, Fast and Slow.* Farrar, Straus and Giroux, New York.

[41] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. doi:10.35542/osf.io/5er8f

[42] Patricia M. King and Karen Strohm Kitchener. 2012. The Reflective Judgment Model: Twenty Years of Research on Epistemic Cognition. In *Personal Epistemology*, Barbara K. Hofer and Paul R. Pintrich (Eds.). Routledge, New York, 37–61.

[43] Chokri Kooli and Nadia Yusuf. 2025. Transforming Educational Assessment: Insights Into the Use of ChatGPT and Large Language Models in Grading. *International Journal of Human–Computer Interaction* 41, 5 (March 2025), 3388–3399. doi:10.1080/10447318.2024.2338330

[44] Markus Krause, Tom Garncarz, JiaoJiao Song, Elizabeth M. Gerber, Brian P. Bailey, and Steven P. Dow. 2017. Critique Style Guide: Improving Crowdsourced Design Feedback with a Natural Language Model. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4627–4639. doi:10.1145/3025453.3025883

[45] Yuheng Li, Mladen Raković, Namrata Srivastava, Xinyu Li, Quanlong Guan, Dragan Gašević, and Guanliang Chen. 2025. Can AI Support Human Grading? Examining Machine Attention and Confidence in Short Answer Scoring. *Computers & Education* 228, C (April 2025), 105244. doi:10.1016/j.compedu.2025.105244

[46] Gregory C Lisby. 2000. *College Student Grade Disputes: Adjudicative vs. Mediative Models of Conflict Resolution.* Ph. D. Dissertation. Georgia State University.

[47] Shusen Liu, Peer-Timo Bremer, Jayaraman J. Thiagarajan, Vivek Srikumar, Bei Wang, Yarden Livnat, and Valerio Pascucci. 2017. Visual Exploration of Semantic Relationships in Neural Word Embeddings. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 553–562.

[48] Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A Token-Level Reference-Free Hallucination Detection Benchmark for Free-Form Text Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 6723–6737. doi:10.18653/v1/2022.acl-long.464

[49] Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. 2024. LLM Comparative Assessment: Zero-Shot NLG Evaluation through Pairwise Comparisons Using Large Language Models. arXiv:2307.07889 [cs] doi:10.48550/arXiv.2307.07889

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

[51] Oscar Luaces, Jorge Díez, and Antonio Bahamonde. 2018. A Peer Assessment Method to Provide Feedback, Consistent Grading and Reduce Students' Burden in Massive Teaching Settings. *Computers & Education* 126 (Nov. 2018), 283–295. doi:10.1016/j.compedu.2018.07.016

[52] Tom Lumley. 2002. Assessment Criteria in a Large-Scale Writing Test: What Do They Really Mean to the Raters? *Language Testing* 19, 3 (July 2002), 246–276. doi:10.1191/0265532202lt230oa

[53] Roberto Martinez-Maldonado. 2019. A Handheld Classroom Dashboard: Teachers' Perspectives on the Use of Real-Time Collaborative Learning Analytics. *International Journal of Computer-Supported Collaborative Learning* 14, 3 (2019), 383–411.

[54] Roberto Martinez-Maldonado, Andrew Clayphan, Kalina Yacef, and Judy Kay. 2015. MTFeedback: Providing Notifications to Enhance Teacher Awareness of Small Group Work in the Classroom. *IEEE Transactions on Learning Technologies* 8, 2 (April 2015), 187–200. doi:10.1109/TLT.2014.2365027

[55] Suzanne McMahon and Ian Jones. 2015. A Comparative Judgement Approach to Teacher Assessment. *Assessment in Education: Principles, Policy & Practice* 22, 3 (July 2015), 368–389. doi:10.1080/0969594X.2014.978839

[56] Emma Mercier. 2016. Teacher Orchestration and Student Learning during Mathematics Activities in a Smart Classroom. *IJSMARTTL* 1, 1 (2016), 33. doi:10.1504/IJSMARTTL.2016.078160

[57] Marcus Messer, Neil C. C. Brown, Michael Kölling, and Miaojing Shi. 2024. Automated Grading and Feedback Tools for Programming Education: A Systematic Review. *ACM Trans. Comput. Educ.* 24, 1 (Feb. 2024), 10:1–10:43. doi:10.1145/3636515

[58] Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to Grade Short Answer Questions Using Semantic Similarity Measures and Dependency Graph Alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (Eds.). Association for Computational Linguistics, Portland, Oregon, USA, 752–762.

[59] Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. Generating Benchmarks for Factuality Evaluation of Language Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 49–66.

[60] Emily T. Ott. 2022. Using Grade Appeals as a Learning Tool. *Teaching in the University* 1 (2022), 1.

[61] Corey Palermo and Margareta Maria Thomson. 2018. Teacher Implementation of Self-Regulated Strategy Development with an Automated Writing Evaluation System: Effects on the Argumentative Writing Performance of Middle School Students. *Contemporary Educational Psychology* 54 (July 2018), 255–270. doi:10.1016/j.cedpsych.2018.07.002

[62] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Hum Factors* 39, 2 (June 1997), 230–253. doi:10.1518/001872097778543886

[63] Les Perelman. 2014. When "the State of the Art" Is Counting Words. *Assessing Writing* 21 (July 2014), 104–111. doi:10.1016/j.asw.2014.05.001

[64] Gustavo Pinto, Isadora Cardoso-Pereira, Danilo Monteiro Ribeiro, Danilo Lucena, Alberto de Souza, and Kiev Gama. 2023. Large Language Models for Education: Grading Open-Ended Questions Using ChatGPT. arXiv:2307.16696 [cs] doi:10.48550/arXiv.2307.16696

[65] Alastair Pollitt and Victoria Crisp. 2004. Could Comparative Judgements of Script Quality Replace Traditional Marking and Improve the Validity of Exam Questions. In *BERA Annual Conference, UMIST Manchester, England*. British Educational Research Association, Manchester, UK, 1–17.

[66] QuillBot. 2025. Free AI-Powered Essay and Paper Checker—QuillBot AI. https://quillbot.com/essay-checker.

[67] Federica Zoe Ricci, Catalina Mari Medina, and Mine Dogucu. 2024. Automated Grading Workflows for Providing Personalized Feedback to Open-Ended Data Science Assignments. arXiv:2309.12924 doi:10.48550/arXiv.2309.12924

[68] Raymond Scupin. 1997. The KJ Method: A Technique for Analyzing Data Derived from Japanese Ethnology. *Human Organization* 56, 2 (1997), 233–237. jstor:44126786

[69] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (April 2020), 495–504. doi:10.1080/10447318.2020.1741118

[70] Valerie J. Shute. 2008. Focus on Formative Feedback. *Review of Educational Research* 78, 1 (March 2008), 153–189. doi:10.3102/0034654307313795

[71] Adele Smolansky, Andrew Cram, Corina Raduescu, Sandris Zeivots, Elaine Huber, and Rene F. Kizilcec. 2023. Educator and Student Perspectives on the Impact of Generative AI on Assessments in Higher Education. In *Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23)*. Association for Computing Machinery, New York, NY, USA, 378–382. doi:10.1145/3573051.3596191

[72] Jacob Steiss, Tamara Tate, Steve Graham, Jazmin Cruz, Michael Hebert, Jiali Wang, Youngsun Moon, Waverly Tseng, Mark Warschauer, and Carol Booth Olson. 2024. Comparing the Quality of Human and ChatGPT Feedback of Students' Writing. *Learning and Instruction* 91 (June 2024), 101894. doi:10.1016/j.learninstruc.2024.101894

[73] Marie Stevenson and Aek Phakiti. 2014. The Effects of Computer-Generated Feedback on the Quality of Writing. *Assessing Writing* 19 (Jan. 2014), 51–65. doi:10.1016/j.asw.2013.11.007

[74] Lu Sun, Aaron Chan, Yun Seo Chang, and Steven P. Dow. 2024. ReviewFlow: Intelligent Scaffolding to Support Academic Peer Reviewing. arXiv:2402.03530 [cs] doi:10.1145/3640543.3645159

[75] Chad C. Tossell, Nathan L. Tenhundfeld, Ali Momen, Katrina Cooley, and Ewart J. De Visser. 2024. Student Perceptions of ChatGPT Use in a College Essay Assignment: Implications for Learning, Grading, and Trust in Artificial Intelligence. *IEEE Transactions on Learning Technologies* 17 (2024), 1069–1081. doi:10.1109/TLT.2024.3355015

[76] Yu-Chia Tseng and Chao Zhang. 2025. The Role of Politeness Strategies in Online Design Feedback Exchange. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3706599.3720100

[77] Peter D. Turney. 2006. Similarity of Semantic Relations. *Computational Linguistics* 32, 3 (2006), 379–416.

[78] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in Judgments Reveal Some Heuristics of Thinking under Uncertainty. *Science* 185, 4157 (Sept. 1974), 1124–1131. doi:10.1126/science.185.4157.1124

[79] Amos Tversky and Daniel Kahneman. 1981. The Framing of Decisions and the Psychology of Choice. *Science* 211, 4481 (Jan. 1981), 453–458. doi:10.1126/science.7455683

[80] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 327:1–327:39. doi:10.1145/3476068

[81] Izia Xiaoxiao Wang, Xihan Wu, Edith Coates, Min Zeng, Jiexin Kuang, Siliang Liu, Mengyang Qiu, and Jungyeul Park. 2024. Neural Automated Writing Evaluation with Corrective Feedback. arXiv:2402.17613 [cs] doi:10.48550/arXiv.2402.17613

[82] D. Watson, L. A. Clark, and A. Tellegen. 1988. Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *J Pers Soc Psychol* 54, 6 (June 1988), 1063–1070. doi:10.1037//0022-3514.54.6.1063

[83] Ping Wei, Xiaosai Wang, and Hui Dong. 2023. The Impact of Automated Writing Evaluation on Second Language Writing Skills of Chinese EFL Learners: A Randomized Controlled Trial. *Front Psychol* 14 (Sept. 2023), 1249991. doi:10.3389/fpsyg.2023.1249991

[84] Joshua Wilson and Amanda Czik. 2016. Automated Essay Evaluation Software in English Language Arts Classrooms: Effects on Teacher Feedback, Student Motivation, and Writing Quality. *Computers & Education* 100 (Sept. 2016), 94–109. doi:10.1016/j.compedu.2016.05.004

[85] Cynthia S. Wiseman. 2012. A Comparison of the Performance of Analytic vs. Holistic Scoring Rubrics to Assess L2 Writing. *International Journal of Language Testing* 2, 1 (March 2012), 59–92.

[86] Kenneth Wolf and Ellen Stevens. 2007. The Role of Rubrics in Advancing and Assessing Student Learning. *Journal of Effective Teaching* 7, 1 (2007), 3–14.

[87] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–22. doi:10.1145/3491102.3517582

[88] Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. From Automation to Augmentation: Large Language Models Elevating Essay Scoring Landscape. arXiv:2401.06431 [cs] doi:10.48550/arXiv.2401.06431

[89] Wenjing Xie, Juxin Niu, Chun Jason Xue, and Nan Guan. 2024. Grade Like a Human: Rethinking Automated Assessment with Large Language Models. arXiv:2405.19694 [cs]

[90] Kexin Bella Yang, LuEttaMae Lawrence, Vanessa Echeverria, Boyuan Guo, Nikol Rummel, and Vincent Aleven. 2021. Surveying Teachers' Preferences and Boundaries Regarding Human-AI Control in Dynamic Pairing of Students for Collaborative Learning. In *Technology-Enhanced Learning for a Free, Safe, and Sustainable World*, Tinne De Laet, Roland Klemke, Carlos Alario-Hoyos, Isabel Hilliger, and Alejandro Ortega-Arranz (Eds.). Vol. 12884. Springer International Publishing, Cham, 260–274. doi:10.1007/978-3-030-86436-1_20

[91] Kexin Bella Yang, Zijing Lu, Vanessa Echeverria, Jonathan Sewall, Luettamae Lawrence, Nikol Rummel, and Vincent Aleven. 2022. Technology Ecosystem for Orchestrating Dynamic Transitions Between Individual and Collaborative AI-Tutored Problem Solving. In *Artificial Intelligence in Education*, Maria Mercedes Rodrigo, Noburu Matsuda, Alexandra I. Cristea, and Vania Dimitrova (Eds.). Vol. 13355. Springer International Publishing, Cham, 673–678. doi:10.1007/978-3-031-11644-5_66

[92] Haneul Yoo, Jieun Han, So-Yeon Ahn, and Alice Oh. 2024. DREsS: Dataset for Rubric-Based Essay Scoring on EFL Writing. arXiv:2402.16733

[93] Audrey Zhang, Yifei Gao, Wannapon Suraworachet, Tanya Nazaretsky, and Mutlu Cukurova. 2025. Evaluating Trust in AI, Human, and Co-Produced Feedback Among Undergraduate Students. arXiv:2504.10961 [cs] doi:10.48550/arXiv.2504.10961

[94] Chao Zhang, Kexin Ju, Peter Bidoshi, Yu-Chun Grace Yen, and Jeffrey M. Rzeszotarski. 2025. Friction: Deciphering Writing Feedback into Writing Revisions through LLM-Assisted Reflection. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–27. doi:10.1145/3706598.3714316

[95] Chao Zhang, Kexin Ju, Zhuolun Han, Yu-Chun Grace Yen, and Jeffrey M. Rzeszotarski. 2025. Synthia: Visually Interpreting and Synthesizing Feedback for Writing Revision. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3746059.3747703

[96] Xuanming Zhang, Anthony Diaz, Zixun Chen, Qingyang Wu, Kun Qian, Erik Voss, and Zhou Yu. 2024. DECOR: Improving Coherence in L2 English Writing with a Novel Benchmark for Incoherence Detection, Reasoning, and Rewriting. arXiv:2406.19650 [cs] doi:10.48550/arXiv.2406.19650