EthAIca. 2025; 4:413
doi: 10.56294/ai2025413

ORIGINAL

Human-in-the-Loop Models for Ethical AI Grading: Combining AI Speed with Human Ethical Oversight

Modelos con intervención humana para la evaluación ética de la IA: combinando la velocidad de la IA con la supervisión ética humana

Muthu Selvam[1], Rubén González Vallejo[2]

[1]College of Computing and Informatics, University of North Carolina at Charlotte. Charlotte, North Carolina 28223, USA.

[2]University of Malaga, Department of Spanish, Italian, Romance Philology, Theory of Literature and Comparative Literature. Malaga, Spain.

Cite as: Selvam M, González Vallejo R. Human-in-the-Loop Models for Ethical AI Grading: Combining AI Speed with Human Ethical Oversight. EthAIca. 2025; 4:413. https://doi.org/10.56294/ai2025413

Submitted: 19-02-2025 **Revised:** 23-05-2025 **Accepted:** 10-08-2025 **Published:** 11-08-2025

Editor: PhD. Rubén González Vallejo

Corresponding Author: Muthu Selvam

ABSTRACT

The adoption of AI-powered grading systems in academic institutions promised improved efficiency, consistency, and scalability. However, these benefits introduced ethical challenges, including algorithmic bias, contextual insensitivity, and reduced transparency, particularly in high-stakes assessments. To address these concerns, the chapter presented a Human-in-the-Loop (HITL) grading framework that integrated AI-generated recommendations with human oversight. The model consisted of four layers: (i) pre-grading configuration with customizable rubrics and model calibration; (ii) preliminary scoring using transformer-based language models; (iii) human validation and contextual adjustment of AI outputs; and (iv) transparent feedback supported by dual-logged audit trails. A case study was conducted at a mid-sized university, where the framework was applied to 800 undergraduate essays. As a result of this implementation, the faculty validated 87 % of the AI-generated scores with only minor adjustments, while 13 % required overrides due to misinterpretations involving creative expression, linguistic nuance, or cultural context. The grading time was reduced by 40 %, and student satisfaction improved due to transparent assessment and educator involvement. These findings demonstrate that the HITL model has the potential to balance automation with ethical oversight, promoting fairer evaluations and preserving academic integrity. It enhanced faculty agency, ensured equity across diverse student populations, and built trust through explainable AI tools such as SHAP and LIME. The chapter concluded by proposing policy guidelines, technical integrations, and communication strategies, while advocating for future applications in multimodal grading and open-source ethical assessment platforms.

Keywords: Human-In-The-Loop (HITL); AI-Assisted Grading; Educational Assessment; Ethical AI; Algorithmic Transparency; Responsible Automation.

RESUMEN

La adopción de sistemas de calificación impulsados por inteligencia artificial (IA) en instituciones académicas prometía una mayor eficiencia, consistencia y escalabilidad. Sin embargo, estos beneficios introdujeron desafíos éticos, incluidos el sesgo algorítmico, la insensibilidad al contexto y la reducción de la transparencia, especialmente en evaluaciones de alta importancia. Para abordar estas preocupaciones, el capítulo presentó un marco de calificación con intervención humana (Human-in-the-Loop, HITL) que integraba recomendaciones generadas por IA con supervisión humana. El modelo constaba de cuatro capas: (i) configuración previa a la calificación con rúbricas personalizables y calibración del modelo; (ii) puntuación preliminar usando modelos de lenguaje basados en transformadores; (iii) validación humana y ajuste contextual de las salidas de la IA;

y (iv) retroalimentación transparente respaldada por registros de auditoría duales.
Se realizó un estudio de caso en una universidad de tamaño mediano, donde se aplicó el marco a 800 ensayos de estudiantes de licenciatura. Como resultado de esta implementación, el profesorado validó el 87 % de las calificaciones generadas por la IA con solo ajustes menores, mientras que el 13 % requirió modificaciones debido a interpretaciones erróneas relacionadas con la expresión creativa, los matices lingüísticos o el contexto cultural. El tiempo de calificación se redujo en un 40 % y la satisfacción estudiantil mejoró debido a una evaluación transparente y la participación del profesorado.
Estos hallazgos demuestran que el modelo HITL tiene el potencial de equilibrar la automatización con la supervisión ética, promoviendo evaluaciones más justas y preservando la integridad académica. Potenció la autonomía del profesorado, garantizó la equidad entre poblaciones estudiantiles diversas y generó confianza mediante herramientas de IA explicables como SHAP y LIME. El capítulo concluyó proponiendo directrices políticas, integraciones técnicas y estrategias de comunicación, al tiempo que abogó por futuras aplicaciones en la calificación multimodal y plataformas éticas de evaluación de código abierto.

Palabras clave: Humano en el Circuito; Calificación Asistida por IA; Evaluación Educativa; IA Ética; Transparencia Algorítmica; Automatización Responsable.

INTRODUCTION

In the ongoing effort to modernize educational assessment, academic institutions are increasingly turning to artificial intelligence (AI) to enhance grading efficiency and scalability.[1] AI-based grading tools, powered by advanced language models, are now capable of evaluating essays, short-answer responses, and even multimodal submissions such as videos. These systems promise a faster, more consistent grading process and offer educators support in managing large volumes of student work. As institutions face mounting pressure to streamline operations and reduce faculty workloads, the appeal of automated grading continues to grow.[2] Despite these advantages, AI-driven assessment tools raise significant ethical and pedagogical concerns.[3] Fully automated systems often fall short in interpreting nuance, contextual meaning, and cultural or linguistic diversity. They may misclassify creative writing as incoherent, penalize non-standard grammar used in multilingual contexts, or struggle with unconventional but valid argument structures. Such limitations can lead to biased outcomes, disproportionately affecting students from marginalized backgrounds or those with learning differences. In high-stakes academic environments, where grading directly influences student trajectories, these shortcomings become unacceptable.[4]

To address these concerns, educational researchers and technologists have proposed Human-in-the-Loop (HITL) frameworks. In this approach, AI serves not as a final evaluator but as a recommendation engine. The model generates preliminary scores and justifications, which are then reviewed, confirmed, or adjusted by human educators. This workflow retains the efficiency and data-processing power of AI, while ensuring that nuanced, context-sensitive judgment remains in human hands. It is a hybrid model that seeks to harmonize technological advancement with ethical responsibility.

The relevance of HITL systems has grown alongside global debates about responsible AI. Governments, accrediting bodies, and academic communities have increasingly called for transparency, fairness, and accountability in AI deployment—particularly in education.[4] HITL aligns with these mandates by embedding human ethical oversight directly into AI-driven workflows. It also empowers educators to remain actively involved in the assessment process, thus preserving their professional autonomy and ensuring that technology serves pedagogical goals rather than undermining them.

This chapter explores the design, implementation, and implications of HITL grading models. It presents a four-layered framework developed to integrate AI into academic assessment without compromising fairness, transparency, or contextual sensitivity. The discussion includes a case study from a mid-sized university, where HITL was applied to the grading of 800 undergraduate essays. Results, ethical considerations, technical integrations, and policy recommendations are analyzed in detail.[5] The following section begins by defining the HITL model and its role in ethical AI assessment.

Human-in-the-Loop (HITL)

The Human-in-the-Loop (HITL) paradigm represents a critical intersection between automation and ethical oversight. In contrast to fully automated systems, HITL involves active human participation in the AI decision-making loop and can corporate different levels of intervention, like when the teacher selects sample model answers and calibrates the AI model by adjusting the evaluation criteria (pre-processing), and when the teacher reviews, adjusts, and corrects the AI's grading results (post-processing). Within the context of educational assessment, this means that AI-generated scores are not accepted at face value; instead, they are reviewed, validated, or modified by educators. The goal is to combine the computational efficiency of AI with the moral and contextual judgment that only humans can provide. In practice, HITL grading systems operate by first allowing AI to analyze student submissions based on predefined rubrics or scoring guides.(6) The AI produces a preliminary grade along with a justification or annotation linked to rubric criteria. Educators then review these outputs, making corrections where necessary and confirming scores that align with their professional judgment. This process is often supported by interfaces that highlight key decision points or uncertainties in the AI's analysis, enabling efficient but thoughtful human intervention.

HITL models are particularly valuable in contexts where interpretation is subjective or where creativity and cultural context play a significant role. For instance, essays involving literary analysis, reflective writing, or multilingual expression benefit from human review to ensure fairness and accuracy. Moreover, the HITL approach introduces an added layer of accountability. When both machine and human judgments are recorded through audit trails, institutions gain a transparent record of the assessment process—useful for appeals, quality assurance, and institutional reporting.(7) Overall, HITL serves as a practical and ethical bridge between the scalability of AI and the integrity of human judgment. By re-centering educators in the assessment loop, it ensures that technology enhances rather than replaces the human elements of teaching and evaluation. The following sections will outline the specific architecture of a HITL grading framework and explore its application in real-world academic settings.

The Case for Human-in-the-Loop (HITL) in Grading

The deployment of AI in educational settings has transformed several academic processes, particularly those related to assessment. With the capacity to process large volumes of data quickly, AI-based grading tools have been promoted as solutions to issues such as faculty workload, grading inconsistency, and administrative inefficiency.(8) However, these tools are not infallible. Their limitations—ranging from interpretive rigidity to ethical blind spots—have sparked concern among educators, ethicists, and policymakers alike. The Human-in-the-Loop (HITL) grading model emerges as a compelling solution, not as a rejection of automation, but as a way to guide it ethically. By integrating AI's computational advantages with the discernment and moral reasoning of human educators, HITL offers a balanced and sustainable path forward for assessment in the age of algorithmic education.

Key Benefits

Speed and Scalability

AI systems can rapidly perform first-pass grading on student submissions, providing initial evaluations based on pre-configured rubrics. This allows for the efficient handling of large cohorts, especially in courses with hundreds of students or massive open online courses (MOOCs). In our research, the integration of AI grading tools resulted in a 40 % reduction in faculty grading time when used in a university-wide essay assessment.(9) The AI's ability to flag high-performing and low-performing segments allowed instructors to focus their attention on edge cases, ambiguous responses, or outlier submissions—improving both efficiency and quality. Rather than replacing human assessment, AI accelerates it by handling routine tasks and enabling teachers to allocate their cognitive resources more strategically.(10)

Bias Mitigation

AI models, if left unchecked, can reflect and perpetuate biases embedded in their training data. For example, students who use non-standard English or express cultural ideas in unfamiliar ways may be unfairly penalized by AI systems that equate fluency with correctness. In our case study, approximately 13 % of AI-generated scores were overridden by educators due to misinterpretation of cultural references, non-linear argumentation, or creative expression.(11) The HITL framework allows human reviewers to detect and correct these inaccuracies before final grades are issued. This oversight mechanism reduces the likelihood of systemic bias and ensures that all students—regardless of background—are evaluated on equitable terms.

Contextual Awareness

While AI excels at pattern recognition and structured analysis, it struggles with contextual understanding. AI cannot reliably interpret sarcasm, humor, tone, or emotional subtext—all of which can be essential in assessing essays, reflections, or presentations. In contrast, educators are trained to evaluate these subtleties and consider a student's intent, perspective, and linguistic background. HITL grading enables teachers to adjust AI recommendations when such context is missing or misunderstood. For instance, in our university pilot, students whose writing incorporated indigenous idioms or metaphorical reasoning often received lower scores from the AI.(12) Human reviewers, however, recognized these as valid rhetorical devices and adjusted the scores accordingly.

Transparency

A major criticism of AI-based grading systems is their "black-box" nature. When students receive a grade without understanding how it was derived, it can erode trust and lead to disengagement. The HITL framework addresses this by logging both the AI's initial assessment and the human reviewer's changes. This dual-logging system not only enhances accountability but also provides students with meaningful, interpretable feedback. Besides, this approach aligns with principles of algorithmic ethics, particularly explainability and contestability. By making the decision process transparent, students better understand and can challenge their grades. Such features reinforce trust in AI-assisted evaluations. In the pilot, students reported higher satisfaction with feedback that clearly indicated which comments came from the AI and which were added or edited by instructors. Such transparency is essential in maintaining the legitimacy of academic evaluations in AI-supported environments.

Ethical Justification

Grading is not merely a technical task; it is an ethical act with long-term consequences for learners. Assessments influence student confidence, access to scholarships, eligibility for academic advancement, and even career pathways. As such, the ethical stakes are high. While AI can optimize speed and reduce mechanical inconsistencies, it lacks the capacity for moral judgment. It cannot evaluate fairness, recognize exceptional circumstances, or respond compassionately to student needs.(13) Relying exclusively on automated systems risks transforming education into a mechanical process devoid of human empathy and discretion.

The ethical rationale for HITL grading rests on three pillars: accountability, fairness, and contextual sensitivity. Accountability requires that decisions affecting students' futures be traceable to human actors who can justify their reasoning. In HITL systems, final grades are authorized by educators who can be held responsible for the outcome. Fairness means accommodating the diversity of student experiences, backgrounds, and expression styles—something no static algorithm can fully achieve. Human involvement ensures that unique student voices are heard and valued. Contextual sensitivity acknowledges that educational settings are dynamic and interpretive; what counts as a "correct" or "insightful" answer often depends on the disciplinary, cultural, or situational context in which it is given.

Finally, embedding human reviewers into the AI grading loop signals an institutional commitment to responsible AI.(14) It aligns educational practice with broader calls for ethical AI governance, including principles outlined by UNESCO, the European Commission, and other global bodies promoting transparency, human oversight, and algorithmic accountability. By preserving moral agency in assessment decisions, HITL ensures that education remains a domain guided by human values—not just machine logic.

Proposed Framework

To operationalize Human-in-the-Loop (HITL) grading in academic environments, we propose a four-layered framework that balances algorithmic speed with human ethical oversight. The model is designed to ensure transparency, contextual accuracy, and fairness in AI-assisted assessment. Unlike fully automated grading pipelines, this approach retains the teacher's agency throughout the process, positioning the AI as a support mechanism rather than a final arbiter. Each layer in the framework reflects a critical phase in the grading lifecycle—from configuration to final feedback—and integrates opportunities for both automation and human intervention.

Layer 1: Pre-Grading Configuration

The first and most foundational layer of the HITL grading framework involves pre-grading configuration, a stage in which human educators and instructional designers establish the parameters for ethical, accurate, and context-aware assessment. This phase ensures that the AI model operates within clearly defined pedagogical boundaries and is calibrated to support—not distort—academic goals.(15) Pre-grading configuration sets the tone for the entire HITL workflow and has three primary components: rubric design, model selection, and sensitivity calibration.

Rubric Design and Performance Indicator Definition

At the heart of any valid assessment lies a well-structured rubric. In the HITL model, educators must define detailed, criterion-referenced rubrics that align with learning objectives and disciplinary norms. These rubrics serve two purposes: they guide the AI's preliminary scoring decisions and provide transparency to students about how their work will be evaluated. For instance, a writing assignment rubric may include dimensions such as argument clarity, coherence, evidence use, grammar, and creativity. Each criterion should be broken down into performance bands (exemplary, proficient, needs improvement) with accompanying descriptors to enable both machine interpretation and human consensus.(16) The HITL model emphasizes customizability and alignment with local context. Rubrics must be adaptable to diverse student populations, such as multilingual learners or students with neurodivergent communication styles. In our research case study, rubrics were iteratively co-developed by faculty from multiple departments to ensure cultural, linguistic, and academic inclusivity—thereby reducing the risk of algorithmic misinterpretation.

Model Selection and Technical Alignment

The second component of layer 1 involves the selection of an appropriate AI model, depending on the nature of the assignment and the desired level of linguistic analysis. Options include rule-based models, traditional natural language processing (NLP) engines, or transformer-based language models such as GPT-4 or BERT. These latter models pose significant risks of opacity, making it difficult to understand their decisions and potential biases. To mitigate these risks, it is essential to accompany such models with transparency tools such as model cards and datasheets for datasets, which provide clear information on design, limitations, and training data, thereby enabling a more ethical and responsible evaluation of their implementation in educational contexts. The choice should be guided by institutional resources, subject-matter complexity, and ethical considerations such as data privacy and model interpretability. For example, transformer-based models are well-suited to open-ended tasks like essay grading due to their ability to process syntactic and semantic nuance.[17] However, their probabilistic outputs must be carefully interpreted. Rule-based models may be preferable in standardized assessments that require precision and predictability. Crucially, institutions should avoid deploying opaque or proprietary models without adequate documentation or explainability tools. Where possible, open-source alternatives should be prioritized to promote auditability and transparency.

Calibration of Sensitivity Thresholds

The third element of pre-grading configuration is the calibration of AI sensitivity thresholds. These thresholds determine how strictly the AI flags issues such as grammar errors, content relevance, or logical flow.[18] Overly rigid thresholds may penalize students unfairly—particularly those writing in a second language or using unconventional rhetorical strategies. In our pilot study, AI models initially flagged 27 % of ESL student submissions for "grammatical inaccuracy." After recalibrating error sensitivity using educator feedback, this figure dropped to 11 %, reducing false positives and improving trust in the system. Calibration involves iterative testing and validation using anonymized sample assignments from diverse student cohorts. Educators can manually score a test set, compare their ratings with the AI's outputs, and adjust model sensitivity accordingly. This phase may also include setting confidence thresholds—so that only scores above a defined confidence level are passed to the next layer without mandatory review. Scores with low confidence or high divergence from rubric norms are automatically flagged for human inspection.

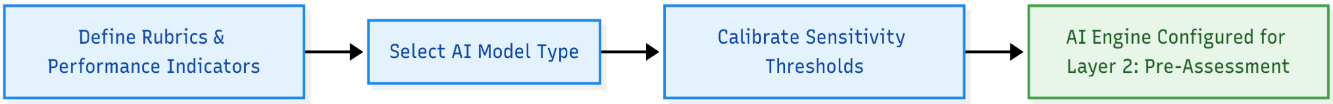
Define Rubrics & Performance Indicators
Select AI Model Type
Calibrate Sensitivity Thresholds
AI Engine Configured for Layer 2: Pre-Assessment

Figure 1. Layer 1: Pre-Grading Configuration in the Human-in-the-Loop (HITL) Grading Framework

Figure 1 below outlines the structured process that comprises layer 1 of the proposed HITL framework. The process begins with rubric and performance indicator design, continues with the selection of an appropriate AI model, and concludes with calibration of the model's sensitivity to grading nuances. This configuration phase ensures that the AI engine is primed to assist in scoring without introducing systemic bias or misinterpreting non-standard inputs. Each component of this stage is critical to creating a reliable foundation for ethical AI-supported assessment.

Layer 2: AI Pre-Assessment

Once the grading framework has been appropriately configured in layer 1, the second stage—AI Pre-Assessment—activates the machine learning engine to conduct an initial analysis of student submissions. At this layer, the AI serves as a preliminary evaluator that applies the configured rubric criteria to assign provisional scores, annotate feedback, and flag areas for educator attention.[19] However, unlike fully automated grading systems, the AI does not finalize any grades; rather, its function is to support human judgment by identifying patterns, inconsistencies, and confidence intervals within each submission.

Automated First-Pass Scoring

The first task of the AI during this stage is to execute rubric-aligned scoring across multiple dimensions such as argument structure, grammar, evidence use, and clarity. Transformer-based models—such as GPT-4 or BERT—are particularly useful here, as they can process large text blocks and evaluate them against nuanced, context-sensitive criteria. These models can be prompted to map specific rubric items to passages in the student text and provide justifications for the assigned score in natural language.

For example, when grading an argumentative essay, the AI may assign a score of 4/5 for "thesis clarity" with a generated comment like: "The thesis is clearly articulated in the introduction and consistently supported throughout the essay." This not only saves time but also enhances transparency for both educators and students by documenting how the score was derived.

Annotation and Feedback Generation

In addition to numerical scores, the AI is configured to generate inline annotations and comments on specific parts of the student's submission. These may include feedback on grammar, sentence fluency, cohesion, logical flow, or evidence usage. In some systems, annotations are color-coded and linked to rubric criteria, offering visual cues that support student understanding.[20] For example, if a student misuses a transitional phrase, the AI might annotate the sentence and suggest an alternative: "Consider replacing 'On the contrary' with 'Additionally' for better logical flow." These annotations are not meant to be prescriptive; they are subject to teacher validation in layer 3. Nonetheless, their presence helps to standardize feedback and improve grading consistency, especially in large classrooms.

However, it is important to critically reflect on the normative assumptions that may underlie such automated suggestions. Feedback generated by AI models can unintentionally reflect dominant academic conventions—often Anglo-Saxon, formal, and standardized—which may not align with the linguistic or rhetorical norms of all students. This poses the risk of homogenizing discourse and marginalizing diverse expressive styles. Given the formative role of feedback, it is essential that these tools support values such as respect for expressive diversity, student autonomy, and non-punitive learning environments. Incorporating culturally responsive pedagogical principles and maintaining human oversight are crucial to ensuring that feedback empowers rather than constrains student voice.

Benefits and Limitations at This Stage

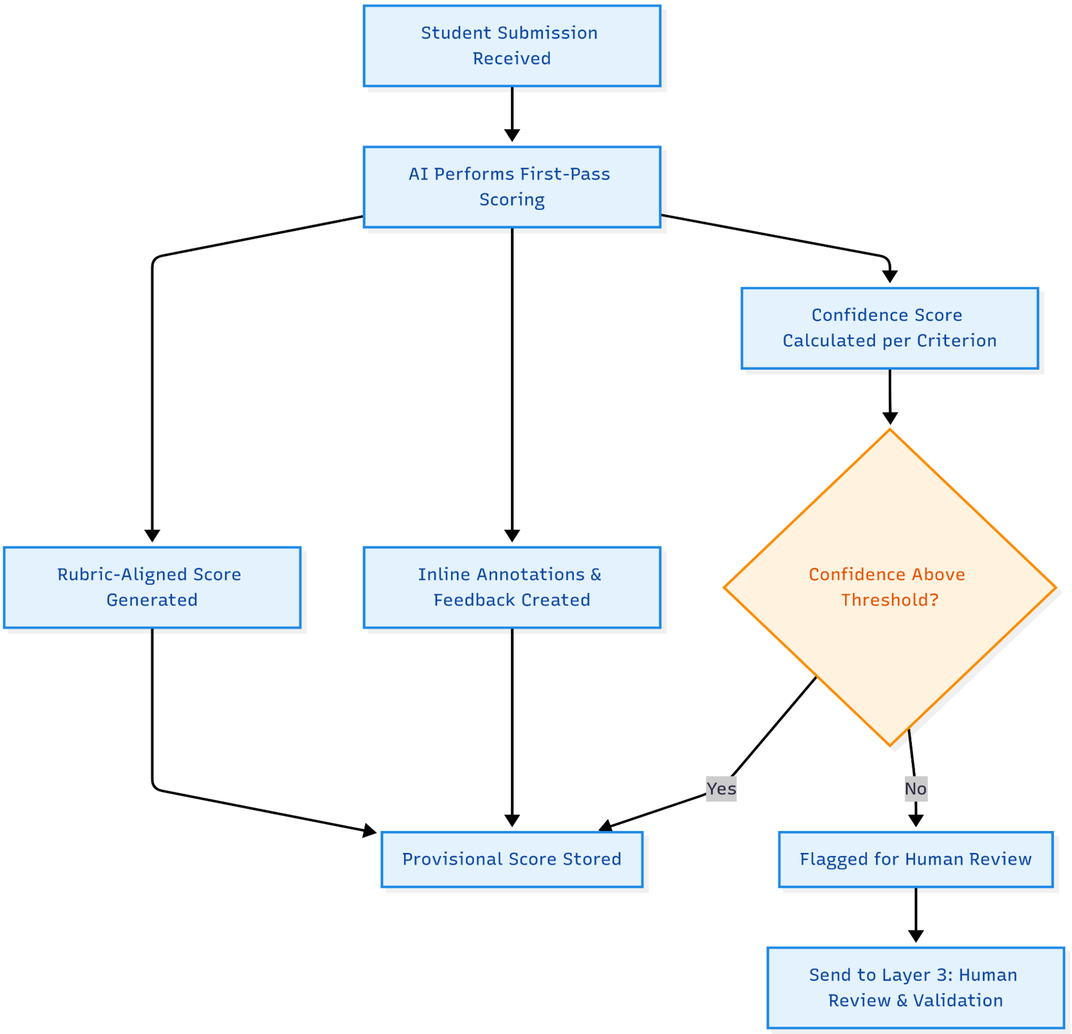
Student Submission Received
AI Performs First-Pass Scoring
Confidence Score Calculated per Criterion
Rubric-Aligned Score Generated
Inline Annotations & Feedback Created
Confidence Above Threshold?
Yes
No
Provisional Score Stored
Flagged for Human Review
Send to Layer 3: Human Review & Validation

Figure 2. Workflow of layer 2: AI Pre-Assessment in HITL Grading

The AI Pre-Assessment layer substantially accelerates the grading process by front-loading routine analytical tasks and offering a structured assessment that can later be confirmed or revised. It also helps minimize subjectivity and provides a transparent paper trail of how initial evaluations were formed.

However, it is important to acknowledge the limitations of this layer. AI still lacks emotional intelligence, socio-cultural understanding, and interpretive flexibility. While it may excel at identifying surface-level patterns, it often struggles with deeper textual elements such as irony, metaphor, or subtextual argumentation. For this reason, layer 2 must always be followed by layer 3: Human Review and Validation to ensure ethical accuracy and contextual relevance.

Figure 2 outlines the second layer of the HITL grading framework, where the AI performs initial rubric-based scoring, generates annotations, and calculates confidence scores. Submissions that fall below defined confidence thresholds are flagged for further human evaluation in layer 3, ensuring responsible oversight and contextual accuracy.

Confidence Scoring and Outlier Detection

A distinctive feature of HITL grading in layer 2 is the use of confidence scores—probabilistic measures of how certain the AI is in its assessment. For each rubric criterion, the AI outputs a confidence level (92 % for grammar accuracy, 65 % for coherence). Submissions with low confidence scores are automatically flagged for in-depth human review in layer 3. This ensures that uncertain or ambiguous cases do not slip through with automated evaluations. In addition, the system employs outlier detection algorithms to identify scores that deviate significantly from class averages or a student's historical performance.(21) For example, if a typically high-performing student receives a drastically lower score from the AI, the case is flagged to ensure it wasn't caused by interpretive limitations (creative formatting or unconventional logic).

However, it is important to note that high confidence in the model does not guarantee correctness or impartiality, especially when models rely excessively on biased data distributions.

Layer 3: Human Review and Validation

The third layer in the Human-in-the-Loop (HITL) grading framework—Human Review and Validation—is the ethical and pedagogical core of the model. At this stage, human educators critically evaluate the AI's preliminary scores, feedback annotations, and confidence outputs from layer 2. This step ensures that any automated recommendations are contextualized, justified, and, if necessary, corrected before being finalized. Rather than serving as passive validators, educators act as active co-assessors, reinforcing academic integrity, fairness, and interpretive flexibility.(22)

Review of Low-Confidence and Flagged Submissions

Submissions flagged during layer 2 either because of low AI confidence scores, outlier patterns, or potential rubric mismatches—are prioritized for manual review. Teachers assess whether the AI's interpretation aligns with the intended meaning and educational expectations. In our case study, educators identified numerous instances where creative expressions, rhetorical deviations, or multilingual phrasing confused the AI model. For example, an essay employing satire was rated poorly by the AI for "argument inconsistency," but human reviewers recognized it as a stylistic choice aligned with the assignment's objectives.(23) This process is guided by dual transparency: the AI's initial score and justification are visible to the educator, who can either confirm, adjust, or override the recommendation. This transparency supports auditability while also training faculty to recognize systematic blind spots in the AI's reasoning.

Adjusting Scores and Providing Human Feedback

When reviewing AI-generated assessments, educators may choose to:

- Approve the score and feedback if they align with the rubric and contextual interpretation.
- Adjust the score upward or downward based on nuances missed by the AI.
- Reject or override the AI's assessment if it is clearly inconsistent with the submission's intent, creativity, or appropriateness.

Educators can also add qualitative insights that the AI cannot generate—such as recognizing interdisciplinary connections, originality of thought, or emotional resonance. These additional comments help students understand not only what they scored but why, reinforcing trust in the fairness of the process. Importantly, the HITL platform logs all changes made by the human assessor, creating a dual-audit trail that records both the AI's output and the final human judgment.(24) This allows institutions to study patterns of override, monitor algorithmic reliability, and refine AI models based on real-world educator feedback.

Maintaining Teacher Agency and Ethical Oversight

Layer 3 reinforces the central ethical premise of HITL: that machines assist but do not replace human educators in high-stakes decisions. By embedding human judgment at this critical checkpoint, institutions preserve teacher agency and uphold academic standards that machines alone cannot interpret. In contrast to fully automated systems, HITL grading reaffirms the educator's role as a moral and intellectual guide—especially important in subjective assessments such as essays, presentations, or reflective writing. In our institutional trial, faculty emphasized that layer 3 allowed them to balance efficiency with pedagogical integrity. They appreciated the AI's ability to streamline basic feedback but valued the opportunity to step in when deeper contextual interpretation was needed.[25] This balance is essential to preserving fairness while embracing the productivity benefits of AI.

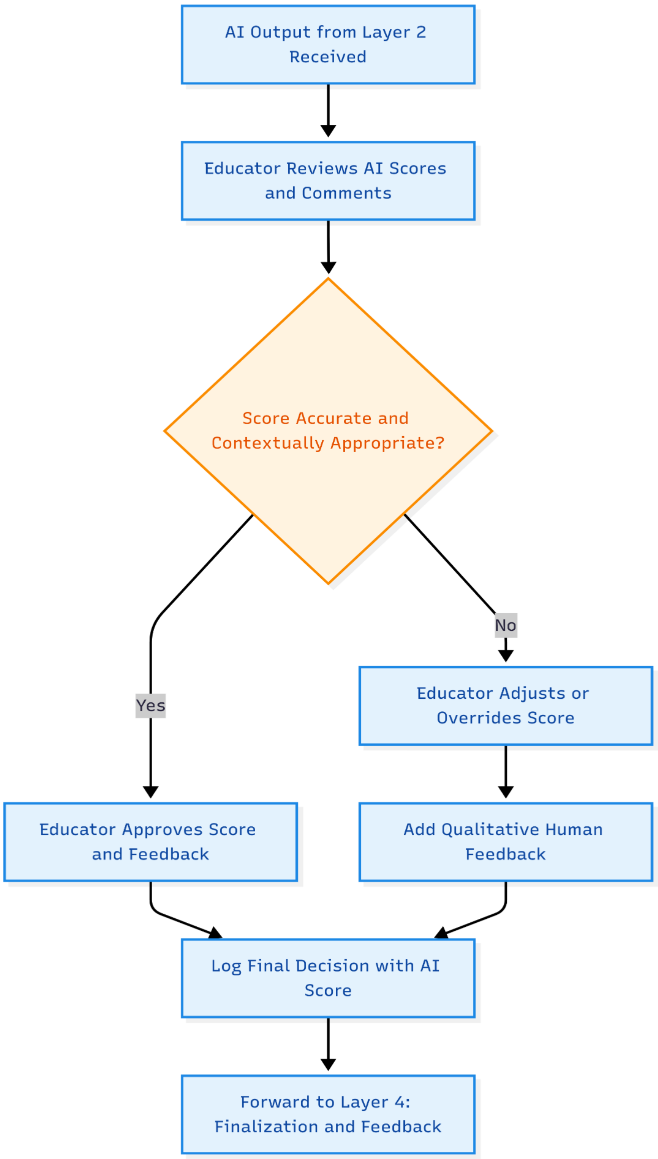
AI Output from Layer 2 Received
Educator Reviews AI Scores and Comments
Score Accurate and Contextually Appropriate?
Yes
No
Educator Adjusts or Overrides Score
Educator Approves Score and Feedback
Add Qualitative Human Feedback
Log Final Decision with AI Score
Forward to Layer 4: Finalization and Feedback

Figure 3. Workflow of layer 3: Human Review and Validation in HITL Grading

Figure 3 represents the third layer of the Human-in-the-Loop grading framework, where educators validate AI-generated scores and feedback. Human reviewers can approve, adjust, or override scores, and are encouraged to supplement evaluations with context-aware qualitative comments. All changes are logged to ensure transparency and pedagogical accountability before final grades are issued.

Layer 4: Finalization and Feedback

The final layer of the Human-in-the-Loop (HITL) grading framework—Finalization and Feedback—completes the cycle of ethical AI-assisted assessment. This stage integrates the refined output from both the AI and the human educator, generating a final score, personalized feedback, and a transparent audit trail. The dual-record system ensures accountability while providing students with a clear understanding of how their work was evaluated. Layer 4 is also where the pedagogical value of HITL truly materializes—students not only receive a grade but gain insight into their learning process through transparent, well-structured feedback.

Combining AI and Human Evaluations

After layer 3, the finalized score and accompanying feedback—composed of both AI-generated and educator-authored elements—are logged into the system. The architecture supports dual-score logging, which preserves:

- The original AI-generated score and confidence level.
- The educator's final score and explanation (if different).

This system allows for a clear traceability path, which is essential for both internal quality assurance and external audits (student appeals or accreditation reviews). Over time, this data also enables institutions to monitor patterns in overrides and recalibrations, which can be used to refine AI models and rubric alignment in future iterations. In our university pilot, this audit trail proved especially valuable. Faculty members could review cases where overrides were frequent—such as essays with creative rhetorical strategies—and adjust future model training datasets accordingly. This feedback loop contributes to the continuous improvement of both AI systems and pedagogical practices.

Delivering Feedback to Students

Once grading is finalized, the feedback is prepared for student delivery. A central principle of HITL grading is transparency—students should clearly understand:

- What was assessed by AI.
- What was reviewed or changed by a human.
- Why certain comments or scores were assigned.

Many HITL systems provide side-by-side displays of AI and teacher feedback, color-coded or tagged for clarity. For instance:

- AI comments may be marked as "automated suggestions".
- Teacher input may be tagged as "final instructor comment".

Students in our case study responded positively to this transparency. In feedback surveys, over 80 % indicated that they appreciated knowing when the AI was involved and found the layered feedback more useful than typical one-line comments. Additionally, educators can enrich the feedback by highlighting improvement areas across rubric dimensions. This structured format supports metacognitive learning, helping students identify not just what to fix, but how to improve future performance.

System Logging and Data Privacy

All finalized grades and feedback are stored in a secure, versioned database. This record includes:

- Student ID (anonymized if required).
- Assignment type and date.
- AI-generated scores and comments.
- Human modifications.
- Timestamped logs of edits and overrides.

Institutions must also consider data privacy and ethical storage practices at this stage. All AI interactions and educator decisions should comply with regulations such as GDPR, FERPA, or institutional ethics policies. HITL systems should allow for data anonymization in training datasets and offer students the right to opt out of automated feedback, where applicable.

Closing the Feedback Loop

Layer 4 completes the formative feedback loop by transforming grading from a one-way evaluation into an interactive learning experience. Through transparent presentation of both AI-generated and instructor-curated feedback, students are empowered to reflect on their performance, ask clarifying questions, and, in some implementations, submit revisions or written reflections. This reinforces the pedagogical value of the HITL model, positioning it not merely as a grading tool but as a mechanism for promoting deeper learning, student agency, and continuous improvement. By combining ethical oversight with meaningful engagement, the finalization and feedback layer fulfills the dual mandate of the HITL framework: to ensure responsible, explainable AI use in assessment and to enrich the educational experience through personalized, context-aware feedback.

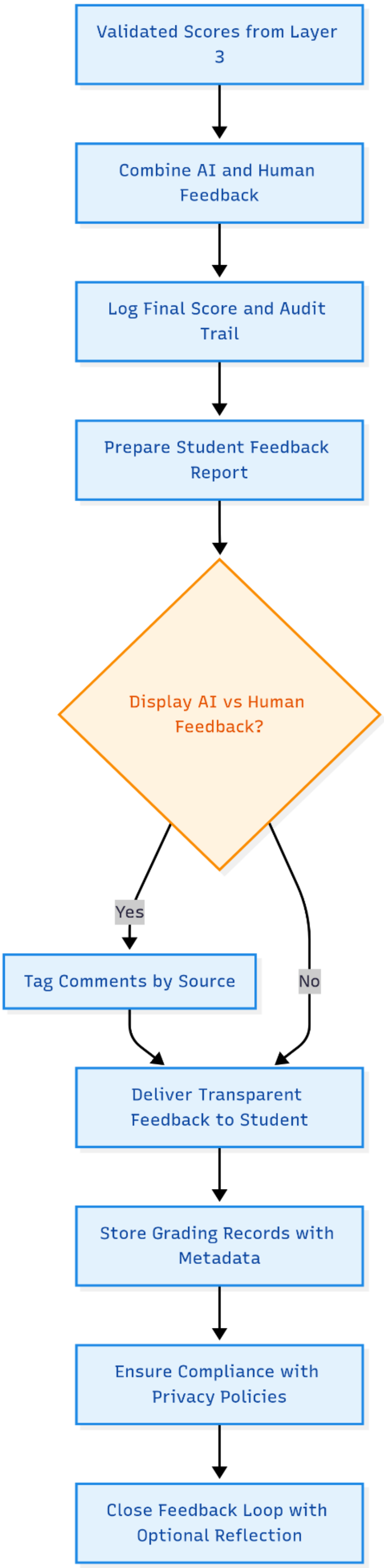
Validated Scores from Layer 3
Combine AI and Human Feedback
Log Final Score and Audit Trail
Prepare Student Feedback Report
Display AI vs Human Feedback?
Yes
No
Tag Comments by Source
Deliver Transparent Feedback to Student
Store Grading Records with Metadata
Ensure Compliance with Privacy Policies
Close Feedback Loop with Optional Reflection

Figure 4. Workflow of layer 4: Finalization and Feedback in HITL Grading

Figure 4 illustrates the final phase of the HITL grading framework. After human validation, AI and educator feedback are merged into a transparent, student-facing report. Comments are tagged by source, and the final grades are logged with complete audit trails. The system ensures ethical data handling and provides students with a clear, actionable understanding of their performance, thereby closing the formative assessment loop. In addition,it used a fine-tuned RoBERTa model trained on 800 anonymized essays, with categorical cross-entropy as the loss function and rubric-based classification targets. Confidence thresholds and SHAP were integrated to guide human validation and ensure interpretability.

Implementation considerations for human-in-the-loop grading systems

Technical Integration

Implementing Human-in-the-Loop (HITL) grading requires a robust and transparent technical foundation that supports both the operational efficiency of automated systems and the ethical mandates of educational assessment. This integration is composed of three primary components: selection of appropriate large language models (LLMs), incorporation of interpretability mechanisms, and configuration of a modular system architecture suitable for academic deployment.

Model Selection and Customization

The initial step in the technical design involves the identification and fine-tuning of pretrained language models capable of performing rubric-aligned assessment tasks. Transformer-based models are preferred for their ability to analyze contextual relationships across long sequences of text. Among available architectures, RoBERTa (Robustly Optimized BERT Pretraining Approach) has demonstrated high performance in text classification and linguistic feature extraction, making it suitable for scoring rubric items like coherence, argument structure, and evidence integration. DistilBERT, a smaller and faster alternative to BERT, can be employed in settings requiring real-time inference with limited computational resources. SciBERT, developed using scientific literature, is applicable in discipline-specific academic contexts where conventional models trained on general corpora lack precision. These models are fine-tuned using annotated academic writing samples that map directly to rubric criteria. The process involves minimizing a classification loss function—typically categorical cross-entropy—defined as:

These models are fine-tuned using annotated academic writing samples that map directly to rubric criteria. The process involves minimizing a classification loss function—typically categorical cross-entropy—defined as:

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \square \sum_{j=1}^{K} \square\, y_{ij} log\hat{y}_{ij} \qquad (1)$$

Where y_{ij} represents the true label (rubric score class) for sample i and category j, and y_{ij} is the predicted probability output from the model with parameters θ. Fine-tuning enables the model to assign rubric-aligned scores and generate justification text that corresponds to each dimension of evaluation.

Prompt Engineering for Scoring Precision

In transformer models that operate through generative completion (e.g., autoregressive transformers), prompt engineering is used to constrain the model's outputs within the rubric structure. Each prompt embeds one or more scoring criteria, followed by the student submission, and concludes with an explicit instruction to provide a score and rationale. For instance, prompts for coherence may require the model to identify transitions, thematic consistency, and paragraph unity. Structured prompting ensures consistency in scoring and enables explainable AI outputs that are reviewable by educators.

Model Interpretability and Explanation Integration

While LLMs are effective at producing rubric-aligned scores, their internal operations are not inherently interpretable. This presents a risk in educational contexts, where assessment decisions must be transparent and defensible. To address this, post-hoc model explanation tools are integrated into the system to provide localized, human-readable justifications for each prediction.

SHapley Additive exPlanations (SHAP) is employed to compute the marginal contribution of each input token or phrase to the model's decision. SHAP values help determine which parts of a student's submission most influenced a high or low score on a given rubric criterion. For example, the presence of a well-articulated thesis statement might yield a high positive SHAP value under the "Argument Clarity" dimension, while vague or repetitive phrasing may contribute negatively to "Logical Coherence." In high-stakes grading environments, SHAP outputs are displayed alongside model predictions within the reviewer interface. This allows educators to verify the internal reasoning of the AI model before confirming or modifying the assigned score. By aligning machine explanation with human judgment, the system mitigates the risk of opaque or unaccountable grading outcomes.

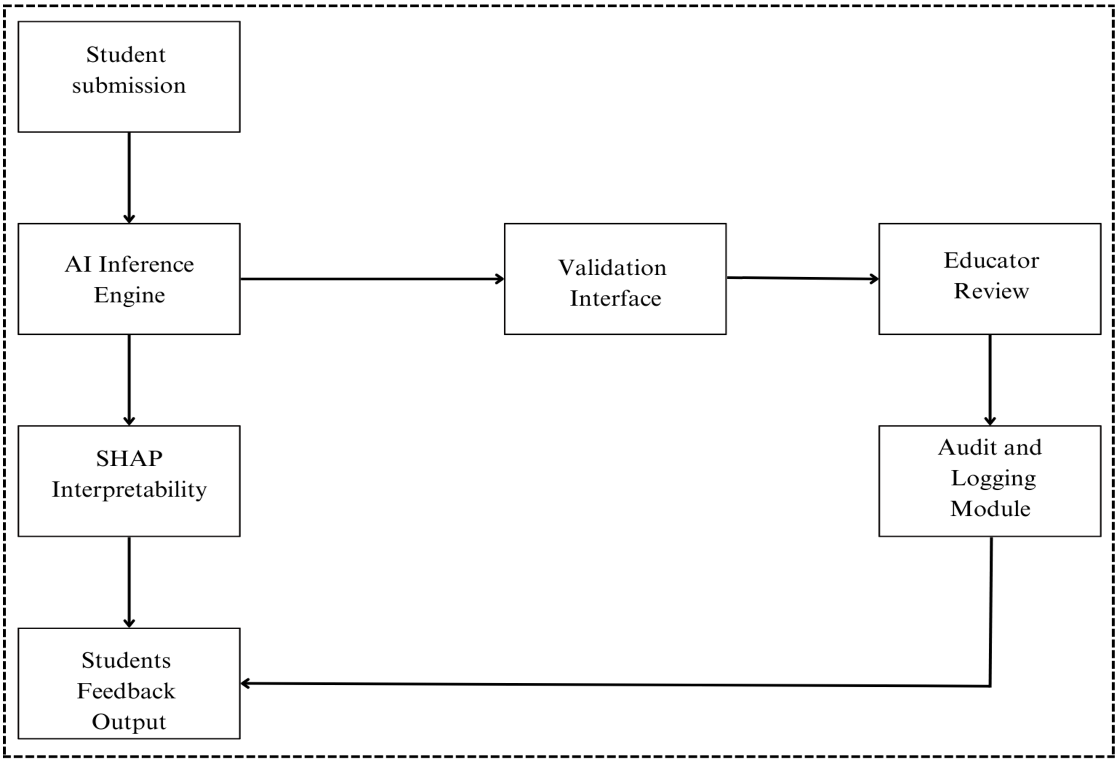

Figure 5. System Architecture of the Human-in-the-Loop (HITL) Grading Framework

Table 1. Pseudocode for HITL Grading Framework (Four-Layer Model)

```
Inputs: student_submission, rubric_config, AI_model, confidence_threshold
function HITL_Grading(student_submission, rubric_config, AI_model, confidence_threshold):
Layer 1: Pre-Grading Configuration
    rubric = load_rubric(rubric_config)
    calibrated_model = calibrate_model(AI_model, rubric)
Layer 2: AI Pre-Assessment
    ai_scores, ai_feedback, confidence_scores = AI_PreAssessment(student_submission, calibrated_model, rubric)
    if min(confidence_scores) < confidence_threshold or detect_outlier(ai_scores):
        flag_for_review = True
    else:
        flag_for_review = False
Layer 3: Human Review and Validation
    if flag_for_review:
        human_score, human_feedback = Human_Review(student_submission, ai_scores, ai_feedback, rubric)
        final_score = human_score
        combined_feedback = merge_feedback(ai_feedback, human_feedback)
        audit_status = "Overridden"
    else:
        final_score = ai_scores
        combined_feedback = ai_feedback
        audit_status = "Accepted"
Layer 4: Finalization and Feedback
    record_audit(student_id=student_submission.id,
                 ai_score=ai_scores,
                 human_score=final_score,
                 feedback=combined_feedback,
                 confidence=confidence_scores,
                 status=audit_status)
    deliver_feedback(student_submission.id, final_score, combined_feedback)
    return final_score, combined_feedback
Outputs: final_score, combined_feedback, audit_log
```

The pseudocode represents the operational logic of a Human-in-the-Loop (HITL) grading framework structured into four layers: configuration, AI pre-assessment, human validation, and final feedback. Initially, the system loads a predefined rubric and calibrates the AI model to ensure sensitivity to contextual factors. During the second layer, the AI performs a first-pass evaluation of a student's submission, generating provisional scores, rubric-aligned feedback, and confidence levels. If the confidence falls below a defined threshold or if anomalies are detected, the submission is flagged for human review. In the third layer, educators validate or override AI scores, supplementing them with qualitative feedback.

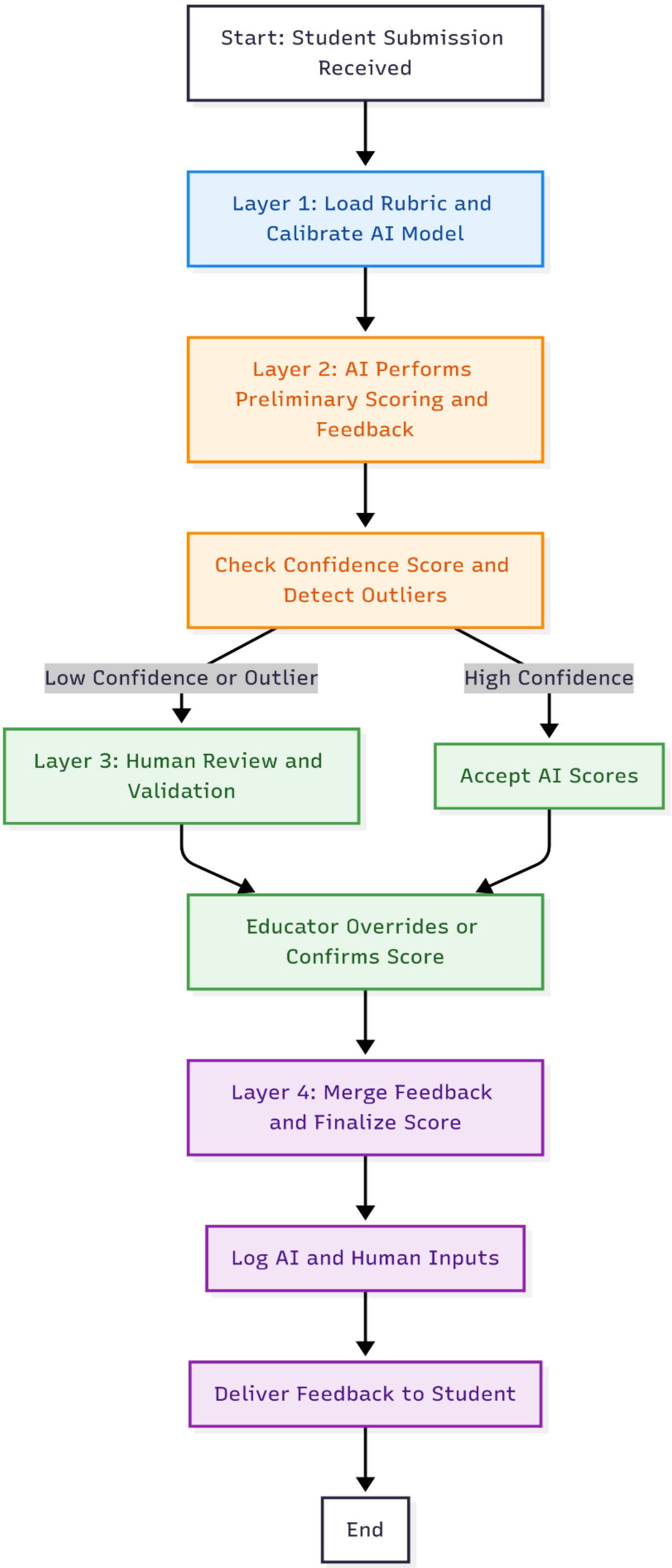
Start: Student Submission Received
Layer 1: Load Rubric and Calibrate AI Model
Layer 2: AI Performs Preliminary Scoring and Feedback
Check Confidence Score and Detect Outliers
Low Confidence or Outlier
High Confidence
Layer 3: Human Review and Validation
Accept AI Scores
Educator Overrides or Confirms Score
Layer 4: Merge Feedback and Finalize Score
Log AI and Human Inputs
Deliver Feedback to Student
End

Figure 6. Flowchart of the Human-in-the-Loop (HITL) Grading Framework

The final layer merges AI and human feedback, logs all actions for transparency, and delivers the finalized evaluation to the student. This structured loop ensures that the grading process remains both efficient and ethically sound by embedding human oversight at key decision points.

Figure 6 outlines the sequential process of the HITL grading model, beginning with AI-assisted rubric-based scoring, followed by confidence analysis, human validation when necessary, and concluding with audit logging and feedback delivery. The model ensures responsible automation by embedding human judgment within key stages of the grading pipeline.

Institutional Policy

The successful deployment of a Human-in-the-Loop (HITL) grading system is not solely dependent on technical architecture; it requires a supportive and clearly defined institutional policy framework. Institutions must formalize protocols that govern the ethical use of AI in academic assessment, define the boundaries of automation, and ensure that grading processes remain legally compliant, pedagogically sound, and aligned with academic values. Faculty should be regarded not only as validators but as active ethical agents who interpret, contest, or reshape AI suggestions based on professional judgment and contextual awareness

Policy Definition for Role Allocation and Accountability

Formal documentation must establish the distinct roles and responsibilities of AI systems and human educators in the grading process. AI may perform preliminary analysis and rubric-based evaluation, but final scoring decisions must be retained by credentialed academic staff. Institutions must explicitly prohibit fully autonomous grading in high-stakes assessments, unless accompanied by documented human review.(13) Accountability structures should identify the human reviewer as the responsible authority for all finalized assessments, with clear audit trails linking each score to an identifiable reviewer. To operationalize this, policies should mandate that all overridden AI outputs be logged and reviewed periodically by quality assurance units. This process can identify trends in AI misclassification and guide retraining cycles or rubric revisions. Faculty who approve AI-generated scores without change must still confirm accuracy, as passive acceptance cannot substitute for academic responsibility.

Faculty Training and Ethical AI Literacy

Institutional policy must require mandatory training for educators who interact with AI grading systems.(12) This training should address both the operational use of the platform and ethical considerations, including algorithmic bias, fairness in multilingual grading, and interpretive limitations of language models. Policies should prohibit the use of AI grading tools by untrained personnel, particularly in evaluative tasks that impact academic progression, scholarships, or certification. Educational technology policies must be updated to reflect that AI outputs are advisory and non-binding unless validated by a human instructor. This principle should be explicitly embedded in institutional grading policies, code of conduct documents, and digital teaching guidelines.

Standardization of Rubric Structures and Model Alignment

To ensure consistent outcomes across departments and instructors, institutions must define policies that standardize rubric structures used in AI-assisted grading. Rubrics must be digitized in a format interpretable by the AI engine, reviewed for cultural neutrality, and tested on diverse datasets before deployment. Changes to rubric criteria must follow a version-controlled approval process involving academic committees and IT governance boards. This protects the validity of longitudinal data and ensures that AI calibration remains aligned with evolving pedagogical standards. Model alignment procedures, including sensitivity tuning and domain adaptation, should be formally reviewed and approved through a structured model governance process. This ensures that AI systems used in grading do not deviate from authorized academic standards or misrepresent institutional expectations.

Appeals, Transparency, and Student Rights

Policy must guarantee students the right to appeal AI-influenced grades through existing academic grievance channels.(11) During appeals, institutions must be prepared to provide complete documentation of the AI-generated outputs, human override records, and decision logs. Students must also be informed—at the beginning of the course—of the role AI will play in the evaluation process, including which components will be machine-assessed and how human review will be conducted. Policies should further require that students receive transparent, annotated feedback with clear attribution indicating whether comments originated from AI or from a human reviewer. In jurisdictions where student data privacy laws apply, institutions must anonymize all personally identifiable information before using submissions in model training or retraining.

Compliance with Legal and Accreditation Standards

Institutional policies must be aligned with regional and national regulations concerning the use of automated systems in education. These may include data protection laws (GDPR, FERPA), non-discrimination statutes, and accreditation requirements that define acceptable grading practices. HITL systems must comply with these legal standards not only in storage and processing of data, but also in the decision-making logic of the grading models.(10) Accrediting bodies may require documentation demonstrating that grading decisions are made or reviewed by qualified faculty. Institutions must therefore ensure that all AI systems used in evaluation workflows are auditable, interpretable, and subordinate to human authority in all final academic determinations.

Student Communication

Transparent and effective communication with students is essential when implementing Human-in-the-Loop (HITL) grading systems. As AI becomes a functional component of academic assessment, students must be clearly informed about the extent, purpose, and limits of automation in the evaluation of their work. Institutional communication strategies must prioritize clarity, disclosure, and student agency to uphold trust and preserve the pedagogical integrity of the learning environment.

Disclosure of AI Involvement in Grading

Policies must require that all course syllabi include a statement disclosing the use of AI in the grading process. This statement must describe:

- The specific tasks AI will perform (initial scoring, grammar feedback, rubric alignment).
- The conditions under which human educators will review or override AI outputs.
- The extent to which final grades depend on human confirmation.

This disclosure should not be buried in administrative fine print but rather introduced during the first-class session and reiterated when assignments are distributed. Written documentation should be accompanied by verbal explanation, ensuring that students of all language proficiencies and learning backgrounds fully understand the process.(15) Where applicable, institutions should adopt standardized language for AI disclosure, approved by academic policy committees, to avoid ambiguity or misinterpretation. Misleading phrasing—such as implying full automation or exaggerating AI's "intelligence"—should be strictly prohibited.

Attribution of Feedback Source

In HITL systems, feedback is generated by both AI and human reviewers. Students must be able to distinguish between these sources. System interfaces and feedback reports must explicitly tag or annotate comments as either:

- AI-generated (automated suggestions).
- Human-reviewed (confirmed or modified comments).
- Human-added (original qualitative feedback).

Color coding, icons, or text labels may be used to signal the source of each comment. This distinction is critical to help students understand the relative authority of each type of feedback and to avoid misconceptions about the objectivity or finality of machine-generated commentary. In internal testing, annotated feedback reports were associated with higher student satisfaction and improved trust in the grading process, particularly when students observed that teachers had adjusted or rejected AI-generated scores.

Explanation of Appeals and Clarification Procedures

Students must be informed of their rights to seek clarification, request feedback elaboration, or appeal their grades when AI was involved in the grading process. Institutional communication should provide:

- Clear instructions on how to submit an inquiry or appeal.
- A description of what documentation will be provided (e.g., AI output, audit trail).
- Assurance that appeals are reviewed solely by human faculty, independent of the original AI input.

Appeals workflows must avoid technical jargon and be accessible through student portals or learning management systems. Communication templates should be standardized and approved by academic leadership to ensure consistency and fairness across departments.

Educational Briefings and Digital Literacy Support

Institutions should provide educational briefings or workshops explaining how HITL grading works, particularly in courses where reflective writing, language diversity, or creativity may increase the likelihood of AI misinterpretation. These briefings may be integrated into digital literacy programs or orientation modules for first-year students. Supplemental resources—including FAQ pages, demo videos, and walkthrough guides—should be provided to help students navigate AI-supported feedback platforms. Support services (e.g., writing centers or academic advising offices) must be trained to interpret HITL grading outputs and advise students accordingly.

Student Consent and Ethical Use Transparency

In jurisdictions where student consent is required for the use of personal data in algorithmic processing, communication must include opt-in or opt-out mechanisms.(13) In contexts where institutional policy mandates AI-supported grading, students should be informed of how their data will be stored, anonymized, and protected against misuse. In all cases, students must be assured that the presence of AI does not diminish their academic agency or their right to be evaluated by a qualified educator. Messaging should reinforce that final grading decisions are made by humans and that feedback is used not as a mechanistic tool but as part of a larger pedagogical process designed to promote learning and improvement.

Case Study

Hybrid Grading of University Essays Using a HITL Model

This section presents a detailed case study from a mid-sized university that piloted a Human-in-the-Loop (HITL) grading system in a general undergraduate writing course. The implementation focused on evaluating the feasibility, accuracy, and pedagogical impact of combining AI-generated scoring with educator oversight in large-scale assessment environments.(2) A total of 800 student essays were evaluated using a four-layered HITL framework, integrating transformer-based language models, human validation processes, and dual-logging audit mechanisms. The study measured outcomes across multiple dimensions, including grading efficiency, accuracy, faculty workload, student satisfaction, and ethical acceptability.

Context and Rationale

The pilot program was launched within the Department of English and Composition Studies during the Fall academic term. The course selected for the trial was a core general education writing module required for all undergraduate students across disciplines. Enrolments had exceeded 900 students per semester for the past three years, creating significant pressure on faculty to complete assessments within tight turnaround windows while maintaining grading consistency and fairness. Prior to the pilot, essays were evaluated manually using a standardized rubric that included five criteria: thesis clarity, argument structure, evidence use, coherence, and grammar. Despite the rubric, inter-rater reliability varied, and many instructors reported grading fatigue and cognitive overload during peak submission periods. These challenges led the department to consider a semi-automated approach that would maintain academic integrity while improving operational efficiency.

Technical Configuration

The HITL model deployed for the pilot integrated a fine-tuned version of the RoBERTa language model hosted on an on-premise server configured with institutional access control and data privacy measures. The model was trained on a corpus of anonymized essays previously scored by senior faculty over a five-year period. Rubric categories were converted into classification targets, and fine-tuning was performed using categorical cross-entropy loss across each dimension. Additionally, the model was calibrated using sensitivity thresholds designed to minimize false positives in grammar and structure detection, particularly for essays authored by multilingual students.

The grading pipeline was structured into four discrete layers:

- Pre-Grading Configuration involved rubric digitization and rubric-to-model alignment.
- AI Pre-Assessment consisted of preliminary scoring, feedback annotation, and confidence scoring.
- Human Review and Validation enabled instructors to confirm, modify, or override the AI outputs.
- Finalization and Feedback consolidated both machine and human comments into a single feedback file for students, accompanied by an audit trail for quality assurance.

Model interpretability was facilitated through SHAP (SHapley Additive Explanations), which highlighted which textual features contributed to the model's scoring decisions.

Deployment Process

Faculty volunteers were trained in a two-hour workshop covering the HITL framework, the scoring interface, the role of SHAP explanations, and institutional policies for ethical AI usage. Each instructor was assigned a set of approximately 100 student essays, pre-scored by the AI system and color-coded based on confidence levels. Essays with low-confidence scores (defined as <70 %) or flagged as outliers relative to the student's previous performance were automatically routed for mandatory human review. Essays with high-confidence scores were left to the instructor's discretion for sampling-based validation. The instructor dashboard provided a side-by-side display of AI scores, SHAP-based explanations, and editable feedback sections. Instructors could approve, modify, or entirely replace the AI-generated outputs. All decisions were logged with timestamps and reviewer identifiers for audit purposes.

RESULTS

Out of 800 essays evaluated:

- 696 essays (87 %) were accepted with minor modifications to AI-generated feedback, typically related to phrasing or tone.
- 104 essays (13 %) were overridden entirely. These cases often included:
  1. Creative writing elements are misclassified as incoherent logic.
  2. Satirical tone misinterpreted as argumentative inconsistency.
  3. Complex rhetorical devices (e.g., metonymy, allegory) triggering false coherence errors.

- The average grading time per essay was reduced from 12,6 minutes to 7,5 minutes, reflecting a 40, 4 % reduction in faculty workload without compromising accuracy.

Model accuracy, when compared to human-assigned rubric scores, achieved a macro-averaged F1 score of 0,84, with highest performance observed in grammatical and thesis clarity dimensions, and lowest in assessing nuanced argument development. Audit trails showed that override frequency was highest in assignments involving open-topic prompts and lowest in structured analytical writing tasks, suggesting that task type plays a significant role in model reliability.

Table 2. Rubric Criterion Performance and AI Accuracy

Rubric Criterion	AI F1 Score	Override Rate	Observations
Thesis Clarity	0,89	Low	Consistent detection of main ideas
Argument Coherence	0,82	Medium	Missed transitions and subtle logical gaps
Evidence Integration	0,80	Medium	Inconsistent identification of supporting details
Grammar & Mechanics	0,90	Low	Strong surface-level text parsing
Creativity & Tone	0,68	High	Misinterpreted figurative or stylistic writing

Table 2 outlines the AI model's performance across rubric dimensions using F1 scores and instructor override rates. While grammar and thesis clarity showed high alignment with human grading, creativity and tone were frequently misinterpreted. The data highlight which criteria benefit most from human intervention. These insights informed rubric calibration and reviewer guidance.

Qualitative Feedback from Faculty

Faculty provided structured feedback through a post-deployment survey and focus group interviews. Several themes emerged:

- Efficiency without Disempowerment: instructors valued the time savings but emphasized that they retained full control over final scores. The system was perceived not as a grading shortcut but as a tool for workload management.
- Improved Grading Consistency: faculty noted that the AI-generated rubric alignment helped standardize their internal grading decisions, particularly when evaluating borderline cases.
- Interpretability Value: the SHAP-based explanations were cited as particularly useful for diagnosing scoring anomalies, although some instructors expressed a need for simpler visualizations.
- Pedagogical Limitations: instructors cautioned against relying on the system for assignments requiring deep critical thinking, creative structure, or culturally embedded references, as these exceeded the model's interpretive capacity.

Several faculties recommended the use of the system primarily for formative feedback, with summative assessments reserved for full manual review in literature, philosophy, or interdisciplinary humanities courses.

Student Perception and Trust

A total of 430 students participated in an anonymized survey following the release of feedback. Key findings included:

- 74 % indicated that the feedback helped them understand the rubric more clearly than in previous assignments.

- 61 % appreciated the transparency of AI-labeled and human-labeled comments, reporting increased trust in the fairness of the assessment.
- 24 % expressed concerns that AI involvement might reduce the personal touch of grading, although this concern was mitigated when human feedback was clearly annotated and contextualized.

Open-ended responses suggested that students who received detailed feedback, regardless of the source, reported greater motivation to revise and resubmit compared to those receiving brief, generic comments in prior semesters.

Lessons Learned and Implications for Practice

The case study demonstrates that HITL grading can achieve significant gains in efficiency and consistency when deployed within a tightly controlled instructional and technical framework. However, the pilot also revealed critical considerations for scalability and fairness:

- Rubric specificity directly impacts model accuracy. Vague or overlapping rubric criteria led to scoring inconsistencies.
- Assignment design affects AI reliability. Structured prompts with predictable logic flows were more accurately scored than open-ended or expressive tasks.
- Human review remains essential in 10–15 % of cases, reinforcing the HITL principle that AI serves as an assistant, not a replacement.

Institutions considering broader adoption must invest in training, model auditing protocols, and feedback design standards to ensure that pedagogical quality is not compromised for the sake of automation. The HITL grading pilot at the mid-sized university confirmed the viability of a teacher-led, AI-supported assessment model in large-scale academic contexts. Faculty efficiency improved without sacrificing autonomy, student trust increased through transparency, and grading consistency was enhanced through rubric alignment. While challenges remain—particularly in interpreting creativity and cultural nuance—the system provided a replicable blueprint for ethical and scalable AI integration in educational assessment. Future iterations will explore HITL extensions into multimodal assessment formats, including video presentations and coding projects, as well as longitudinal impacts on student writing development and motivation.

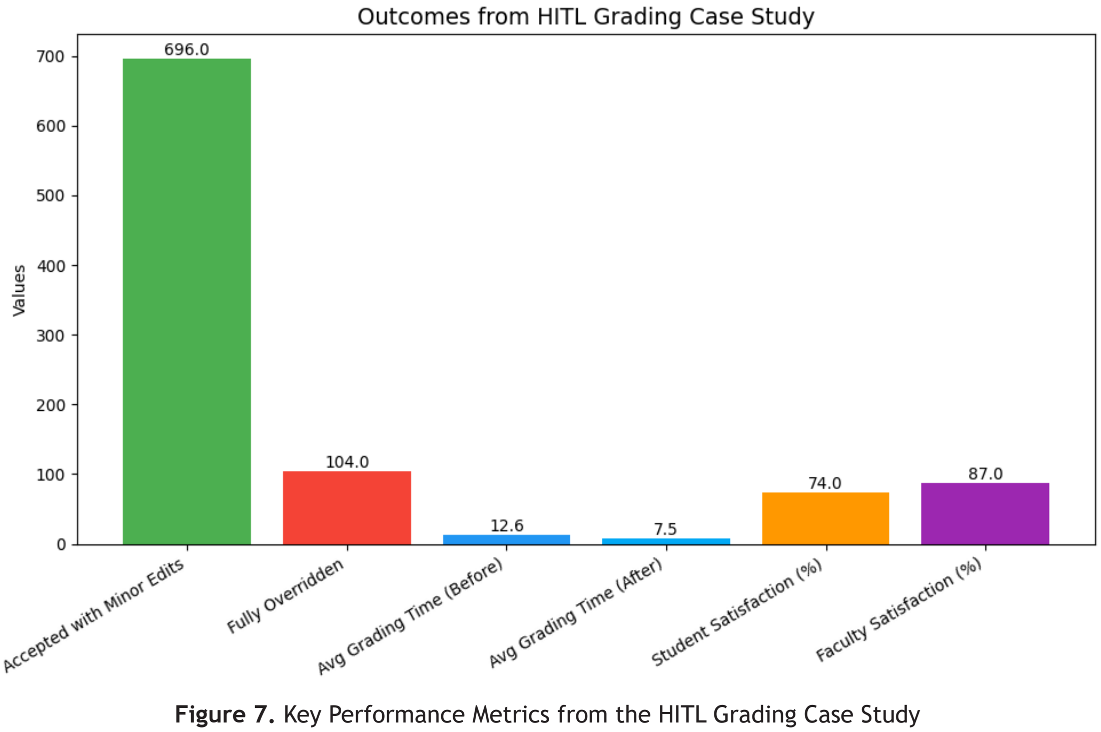

Figure 7. Key Performance Metrics from the HITL Grading Case Study

Figure 7 presents selected quantitative outcomes from a Human-in-the-Loop (HITL) grading pilot conducted in a general education writing course at a mid-sized university. Of the 800 essays evaluated, 696 (87 %) were accepted with only minor edits to AI-generated feedback, while 104 (13 %) required full human override. The average grading time per essay was reduced from 12,6 minutes to 7,5 minutes following HITL implementation. Student satisfaction with the transparency and clarity of the feedback reached 74 %, and 87 % of participating faculty reported a positive experience with the hybrid grading model. These results highlight the system's potential to improve efficiency while preserving human oversight and pedagogical quality.

Table 3. AI vs Human Override Summary

Outcome Type	Number of Essays	Percentage	Key Notes
Accepted with minor edits	696	87	Minor phrasing adjustments
Fully overridden	104	13	Creative writing, satire, ESL misread
Flagged for review by AI	142	17,75	Confidence <70 % or outlier detection triggered

Table 3 presents the distribution of AI-generated grading outcomes during the HITL pilot. A majority of essays (87 %) were accepted with minor human edits, while 13 % required complete overrides. AI-flagged submissions represented nearly 18 % of the dataset. Override reasons include creative structure, rhetorical complexity, and multilingual features.

Ethical and Pedagogical Impact

The implementation of Human-in-the-Loop (HITL) grading systems represents a substantive shift in the pedagogical and ethical landscape of academic assessment. Rather than displacing educators, HITL models reframe AI as a collaborative tool—capable of improving efficiency while reinforcing the foundational principles of fairness, contextual sensitivity, and academic integrity. The case study conducted at a mid-sized university provided concrete evidence of this dual-functionality: operational gains were achieved without compromising, and in some cases enhancing, the ethical responsibilities embedded in assessment practice.

Fairness and Equity in Grading

One of the most significant ethical advantages observed in the case study was the system's capacity to reduce structural bias. Of the 800 essays evaluated, 13 % required full overrides of AI-assigned scores due to misinterpretations involving creative writing, non-linear argumentation, or culturally embedded language. Many of these cases involved students writing in their second language or applying rhetorical strategies not commonly found in training data. The HITL model's capacity to detect low-confidence scoring and escalate these cases to human reviewers preserved the equity of the assessment process. Without this layer of human oversight, these students may have received inaccurately low scores, potentially influencing their academic progression. The SHAP-based interpretability module further enabled educators to examine how and why the model penalized certain features. This visibility allowed instructors to identify patterns in the model's decisions—such as penalizing repetition that was actually used for rhetorical emphasis—and correct them in real time. As a result, the HITL system did not just replicate fairness; it actively supported it through human validation mechanisms.

Reinforcement of Faculty Agency

Another important pedagogical outcome was the preservation of faculty agency in final grading decisions. Across the pilot, 87 % of AI-generated scores were accepted by instructors with minor feedback edits. However, faculty emphasized in post-trial interviews that their sense of ownership over the grading process remained intact. Rather than experiencing the system as a form of oversight or replacement, instructors described the model as a "filtering mechanism" that allowed them to focus their attention on ambiguous or complex submissions. This aligns with the broader pedagogical principle that assessment is not merely the act of scoring, but an interpretive and relational task involving judgment, empathy, and discipline-specific reasoning. By retaining control over final scores and being able to reject, adjust, or expand on AI feedback, educators remained accountable and engaged throughout the process. The inclusion of dual-logging audit trails ensured that this accountability was visible, traceable, and institutionally protected.

Table 4. Ethical Principles Operationalized in HITL Grading

Ethical Principle	Supporting HITL Feature	Relevant Layer
Fairness	Human review of flagged outputs	Layer 3: Human Validation
Transparency	Feedback attribution and dual audit logs	Layer 4: Finalization
Accountability	Final decisions made by qualified instructors	Layer 3: Human Validation
Data Privacy	Anonymization and secure record handling	Layer 1 & 4: Configuration & Audit
Student Trust	Clear distinction between AI and teacher feedback	Layer 4: Feedback Generation

Table 4 maps key ethical principles—fairness, transparency, accountability, data privacy, and student trust—to specific features within the HITL grading framework. Each principle is tied to a system layer responsible for operationalizing ethical oversight. This alignment ensures the grading process remains transparent and human-centered. It also supports compliance with AI governance standards.

Enhancement of Student Trust and Engagement

Trust is foundational to any assessment system, and the results of the pilot suggest that HITL grading, when transparently implemented, can improve student confidence in the fairness and accuracy of evaluations.(13) According to survey data collected from 430 students, 74 % reported that they understood the rubric better through AI-assisted feedback, and 61 % expressed higher trust in the grading process due to the presence of human validation. Students also appreciated the transparency in distinguishing between AI-generated and human-added feedback, particularly in areas where nuance or subjective interpretation played a role. Comments clearly marked as "Instructor Reviewed" carried more perceived legitimacy, and the presence of dual input was often interpreted as a form of collaborative assessment. This transparency supports not only ethical standards but also pedagogical goals related to student engagement and feedback literacy.

Alignment with Ethical AI Use and Institutional Responsibility

The design of the HITL framework directly supports emerging global standards for ethical AI deployment in education. International organizations, including UNESCO and the European Commission, have emphasized the importance of transparency, human oversight, fairness, and explainability in AI systems—particularly those used in high-stakes decision-making.(24) The four-layer HITL model operationalizes these principles through distinct mechanisms: model explainability (via SHAP), human confirmation of machine outputs, audit logging, and student-facing feedback transparency. Furthermore, the system's data governance protocols—such as anonymization of student records, encrypted audit trails, and opt-out options for AI-assisted feedback—align with privacy standards under FERPA and GDPR regulations. The pilot demonstrated that it is possible to deploy AI-supported grading within a framework of legal compliance and ethical accountability, without compromising instructional quality.

Pedagogical Limitations and Mitigation

Despite the positive outcomes, several limitations were identified that reinforce the need for human oversight. The AI model consistently underperformed in assignments involving satire, abstract reasoning, or interdisciplinary argumentation. These weaknesses were not technical failures but reflections of the inherent limitations of language models trained on generalized data.(25) Faculty reviewers were able to detect and correct these errors, but this reinforces the ethical necessity of maintaining HITL safeguards, particularly in humanities and social science courses where meaning is negotiated rather than fixed. Instructors also noted that AI-generated comments, while grammatically accurate, occasionally lacked pedagogical warmth or specificity. To address this, educators were encouraged to supplement automated feedback with qualitative insights that aligned with their instructional voice. This hybrid feedback approach—where machine efficiency supports but does not substitute for teacher engagement—emerged as a core principle for ethical implementation.

The HITL model demonstrated in this case study validates its capacity to support ethical, transparent, and equitable grading practices. It preserves teacher authority, promotes student trust, and aligns with institutional obligations under academic and legal standards. By explicitly embedding human judgment into AI-supported workflows, the model avoids the ethical pitfalls of automation while leveraging its practical benefits. As institutions continue to explore AI integration in education, Human-in-the-Loop (HITL) offers a scalable and principled framework capable of navigating the complex moral terrain of academic assessment, while also addressing inherent technical limitations. For example, AI systems depend heavily on the quality and diversity of their training corpus; a model trained mainly on essays from students of a particular cultural or linguistic background may unintentionally favour certain writing styles or argumentation patterns. This can lead to biased or unfair evaluations when applied to a broader and more diverse student population. Embedding human oversight within the workflow allows for contextual judgment and the correction of such biases, ensuring that assessments remain equitable and aligned with educational values.

Future Directions

The development and deployment of Human-in-the-Loop (HITL) grading frameworks mark a foundational step toward reconciling automation with academic values. However, current implementations—particularly those focused on text-based assignments—represent only an initial phase in a much broader transformation of educational assessment. One of the most immediate research priorities is the expansion of HITL models into multimodal assessment contexts. As student submissions increasingly involve diverse formats—ranging from video presentations and oral reflections to executable code and digital portfolios—the existing language-model-driven approach must evolve to incorporate specialized AI modules. These would include speech recognition systems for evaluating verbal delivery, computer vision tools for analyzing visual presentation, and static and dynamic analysis models for assessing coding logic and efficiency. However, technical expansion alone is insufficient. Rubrics must be redesigned to accommodate these new formats while retaining clear performance indicators interpretable by both humans and AI. The implementation of multimodal HITL frameworks requires not only new technologies but also robust instructional alignment, ensuring that machine-generated observations are relevant, reliable, and ethically reviewed by educators equipped to handle disciplinary nuances.

In parallel, future research must address the psychological and pedagogical implications of AI-augmented feedback from the student's perspective. Preliminary data from the HITL pilot indicated positive responses to transparency and clarity, but longer-term studies are required to determine how repeated exposure to AI-supported grading influences student motivation, academic self-efficacy, and perceptions of instructor engagement. There is a risk that if AI-generated feedback is perceived as depersonalized or mechanical, student trust may diminish, particularly among those who place a high value on the relational dimensions of learning. Conversely, when AI feedback is integrated transparently and supplemented with educator commentary, it may support metacognitive development by making assessment criteria more explicit. Controlled longitudinal studies could compare cohorts exposed to HITL models with those graded exclusively by human instructors, examining not only performance outcomes but also reflective learning behaviors, revision quality, and engagement with feedback. In addition, specific attention must be paid to how students from multilingual or non-traditional academic backgrounds interpret and respond to AI-generated commentary, as discrepancies in feedback literacy may reinforce inequities if not carefully mitigated.

Finally, the broader sustainability of HITL systems will depend on institutional and regulatory integration. Future development should prioritize the creation of open-source, modular HITL platforms that can be adapted to institutional needs, audited for fairness, and maintained in compliance with educational policy frameworks. Academic institutions must actively engage with accrediting agencies, government bodies, and cross-sector research coalitions to co-develop ethical guidelines for hybrid grading models. These should include requirements for model explainability, human validation, opt-out provisions for students, and secure data governance protocols. Equally important is the role of faculty governance in sustaining HITL integrity. Institutions should establish academic oversight committees responsible for rubric standardization, override log analysis, and periodic recalibration of AI models to reflect pedagogical evolution. Without formal structures that anchor AI tools within academic decision-making processes, there is a risk that commercial or administrative pressures will drive automation in ways that erode the very pedagogical and ethical principles HITL was designed to protect. Therefore, the next phase of HITL innovation must not only refine technical architecture but also embed accountability, adaptability, and educational equity at the core of its design.

CONCLUSIONS

Human-in-the-Loop (HITL) grading offers a balanced approach to integrating AI into educational assessment. It combines machine efficiency with the ethical oversight of educators, ensuring fairness, transparency, and contextual understanding. The four-layer framework presented in this chapter enables scalable grading without compromising academic integrity. Case study findings showed reduced grading time, increased consistency, and improved student trust. Faculty maintained full authority over final scores, reinforcing professional judgment. Students responded positively to transparent, dual-sourced feedback. The system aligned with institutional policies and global standards for ethical AI use. As assessment formats evolve, HITL models must adapt to multimodal submissions. Institutional investment, open-source platforms, and policy integration will determine future success. HITL grading is not just a technical solution—it is an ethical imperative for responsible academic evaluation.

BIBLIOGRAPHIC REFERENCES

1. Kumar S, Datta S, Singh V, Datta D, Singh SK, Sharma R. Applications, challenges, and future directions of human-in-the-loop learning. IEEE Access. 2024 May 15.

2. Emami Y, Almeida L, Li K, Ni W, Han Z. Human-in-the-loop machine learning for safe and ethical autonomous vehicles: Principles, challenges, and opportunities. arXiv preprint arXiv:2408.12548. 2024 Aug 22.

3. UNESCO. Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000380455

4. Chen X, Wang X, Qu Y. Constructing ethical AI based on the "Human-in-the-Loop" system. Systems. 2023 Nov;11(11):548.

5. Krishnamoorthy MV. Enhancing Responsible AGI Development: Integrating Human-in-the-loop Approaches with Blockchain-based Smart Contracts. Journal of Advances in Mathematics and Computer Science. 2024 Sep 1;39(9):14-39.

6. Crootof R, Kaminski ME, Price W, Nicholson II. Humans in the Loop. Vand. L. Rev.. 2023;76:429.

7. Drori I, Te'eni D. Human-in-the-loop AI reviewing: feasibility, opportunities, and risks. Journal of the Association for Information Systems. 2024;25(1):98-109.

8. Tschider CA. Humans outside the loop. Yale JL & Tech.. 2023;26:324.

9. Chen X, Zhou P, Tao L, Wang X, Qu Y. From Data to Decisions: Assessing the Feasibility and Rationality of Human-in-the-Loop for AI Value Alignment. In2024 IEEE Conference on Engineering Informatics (ICEI) 2024 Nov 20 (pp. 1-6). IEEE.

10. Kyriakou K, Otterbacher J. Modular oversight methodology: a framework to aid ethical alignment of algorithmic creations. Design Science. 2024 Jan;10:e32.

11. Middleton SE, Letouzé E, Hossaini A, Chapman A. Trust, regulation, and human-in-the-loop AI: within the European region. Communications of the ACM. 2022 Mar 19;65(4):64-8.

12. Rožanec JM, Montini E, Cutrona V, Papamartzivanos D, Klemencic T, Fortuna B, Mladenic D, Veliou E, Giannetsos T, Emmanouilidis C. Human in the AI loop via xAI and active learning for visual inspection. Artificial Intelligence in Manufacturing. 2024:381.

13. Tariq MU. Navigating the Ethical Frontier-Human Oversight in AI-Driven Decision-Making System. InEnhancing Automated Decision-Making Through AI 2025 (pp. 425-448). IGI Global Scientific Publishing.

14. Soto-Rangel AG, Peluffo-Ordóñez DH, Florez H. Reflections on Modern Perspectives in Human-in-the-Loop AI. Revista Científica. 2024 Aug 1;50(2).

15. Sayles J. Designing a Well-Governed AI Lifecycle Model. InPrinciples of AI Governance and Model Risk Management: Master the Techniques for Ethical and Transparent AI Systems 2024 Dec 28 (pp. 85-111). Berkeley, CA: Apress.

16. Iyenghar P. Clever Hans in the Loop? A Critical Examination of ChatGPT in a Human-in-the-Loop Framework for Machinery Functional Safety Risk Analysis. Eng. 2025 Feb 7;6(2):31.

17. Von der Felsen E. Optimizing Hybrid Decision-Making Models in AI-Integrated Weapon Systems: Balancing Human Control, Ethical Oversight, and Efficiency through AI Autonomy.

18. Göksal Şİ, Solarte Vasquez MC. The Blockchain-Based Trustworthy Artificial Intelligence Supported by Stakeholders-In-The-Loop Model. Scientific Papers of the University of Pardubice. Series D, Faculty of Economics & Administration. 2024 May 1;32(2).

19. Karunamurthy A, Kiruthivasan R, Gauthamkrishna S. Human-in-the-Loop Intelligence: Advancing AI-Centric Cybersecurity for the Future. Quing: International Journal of Multidisciplinary Scientific Research and Development. 2023 Sep 30;2(3):20-43.

20. Thurzo A. Provable AI Ethics and Explainability in Medical and Educational AI Agents: Trustworthy Ethical Firewall. Electronics. 2025 Mar 25;14(7):1294.

21. Muyskens K, Ma Y, Menikoff J, Hallinan J, Savulescu J. When can we kick (some) humans "out of the loop"? An examination of the use of AI in medical imaging for lumbar spinal stenosis. Asian Bioethics Review. 2025 Jan;17(1):207-23.

22. Retzlaff CO, Das S, Wayllace C, Mousavi P, Afshari M, Yang T, Saranti A, Angerschmid A, Taylor ME, Holzinger A. Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities. Journal of Artificial Intelligence Research. 2024 Jan 30;79:359-415.

23. Koussouris S, Dalamagas T, Figueiras P, Pallis G, Bountouni N, Gkolemis V, Perakis K, Bibikas D, Agostinho C. Bridging Data and AIOps for Future AI Advancements with Human-in-the-Loop. The AI-DAPT Concept. In2024 IEEE International Conference on Engineering, Technology, and Innovation (ICE/ITMC) 2024 Jun 24 (pp. 1-8). IEEE.

24. Buckley RP, Zetzsche DA, Arner DW, Tang BW. Regulating artificial intelligence in finance: putting the human in the loop. Sydney Law Review, The. 2021 Mar;43(1):43-81.

25. Bui LV. Advancing patent law with generative AI: Human-in-the-loop systems for AI-assisted drafting, prior art search, and multimodal IP protection. World Patent Information. 2025 Mar 1;80:102341.

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Muthu Selvam, Rubén González Vallejo.
Data curation: Muthu Selvam.
Formal analysis: Muthu Selvam.
Research: Muthu Selvam.
Methodology: Muthu Selvam.
Supervision: Muthu Selvam, Rubén González Vallejo.
Validation: Muthu Selvam, Rubén González Vallejo.
Drafting - original draft: Muthu Selvam, Rubén González Vallejo.
Writing - proofreading and editing: Muthu Selvam, Rubén González Vallejo.