Trust-Aware Selective Grading with Retrieval-Grounded Multi-Agent Reflexion

Anonymous ACL submission

Abstract

Large Language Models (LLMs) can be used for automated large-scale evaluation but suffer from hallucinated feedback, fabricated supporting evidence, poor rubric adherence, and unreliable grading under uncertain conditions. To address these problems, we propose a trust-aware automated grading framework that combines a Multi-Agent Reflexion architecture with a Two-Stage RAG pipeline. Our proposed solution parses documents from multiple media formats using a hybrid heuristic OCR process and grounds evaluations using trusted course material through retrieval-augmented generation. The evaluation is performed through an Actor-Critic paradigm where the Critic agent independently validates feedback for logical consistency and factual correctness, while failed evaluations are iteratively improved using a reflexion component. We further introduce a trustworthiness estimation mechanism that combines retrieval confidence, critic confidence, and reflexion burden to determine whether evaluations should be accepted autonomously or escalated for human review. We empirically analyze our framework on submissions from three academic subjects (Distributed Computing, Humanities, and Marketing). As a result, we achieve a hallucinated evidence rate of only 2.9%, while baseline zero-shot LLMs generate feedback with rates as high as 14.0%. Furthermore, the framework achieves a low MAE of 1.9 with a high correlation coefficient of 0.987 while effectively identifying uncertain evaluations unsuitable for fully automated grading. At a trustworthiness threshold of $\phi = 0.8$, the framework autonomously grades 76% of submissions with substantially lower grading error, while uncertain evaluations are selectively escalated for human review.

1 Introduction

The quick development of LLMs has generated a lot of interest in automated assessment in education because of its ability to scale personalized feedback and alleviate the workload of teachers. (Xu et al., 2024; Saputra et al., 2024) While traditional automated grading systems were using rigid approaches such as keyword matching or regular expressions, which do not account for the context of student answers, the problem has been overcome by current approaches with the help of deep contextual understanding (Vaswani et al., 2017; Brown et al., 2020; OpenAI, 2023). Yet, while addressing these problems, there appear several important issues that come along with zero-shot or minimally prompted LLMs in a high-stakes scenario (Gardner et al., 2021; Uyar and Büyükahıska, 2025). Firstly, models are prone to generating imaginary facts, making up fictive quotes from students' papers, and deviating from the defined rubric (Xu et al., 2024; Uyar and Büyükahıska, 2025; Lin et al., 2022; Wang et al., 2024).

Apart from the above issues related to facts generation and rubric deviations, another critical limitation of current automated grading systems is the lack of self-awareness. Today's systems generate evaluations even in cases when the found evidence was incorrect, the process involved re-evaluations because of the error rate, and the system itself was uncertain about the outcome. When it comes to high-stakes academic assessment, the ability of models to recognize cases of their ineffectiveness becomes crucial.

To overcome the discrepancy between the flexibility of the generator and the need for a trustworthy academic assessment, we advocate a trust-aware framework for automated grading that employs retrieval grounding, multi-agent verification, and selective automation under uncertain circumstances. Contrary to existing one-pass LLM grading solutions (Wu et al., 2023; Shinn et al., 2023), our framework is grounded on trusted course material and makes explicit the point at which automated grading may be considered reliable.

Mixed-format student answers are processed

through a hybrid document parser, while a Two-Stage Retrieval-Augmented Generation (RAG) module (Lewis et al., 2020; Izacard and Grave, 2020) is employed to find reliable contextual information grounded on evidence from the corresponding course materials. Evaluations are generated via an Actor-Critic Reflexion loop (Wu et al., 2023; Shinn et al., 2023) wherein the Actor produces a preliminary grade, Critic verifies its consistency with facts and rubrics, and the loop continues whenever inconsistencies are discovered.

While previous studies have separately investigated retrieval-augmented generation and multi-agent reasoning, existing automated grading methods always yield the evaluation output irrespective of uncertainty. Our proposed solution adopts a trust-aware framework for selective grading that is based on retrieval grounding, Actor-Critic Reflexion, and trustworthy evaluation detection. Moreover, we propose evidence-grounded score generation along with quantitative measures of hallucination and escalation for evaluating automated grading under uncertain conditions.

The primary contributions of the paper:

1. **Trust-Aware Selective Grading Framework:** We propose a trust-aware automated grading architecture that combines retrieval grounding, multi-agent verification, and selective escalation to determine when grading can be performed autonomously and when human review is required.
2. **Multi-Signal Trustworthiness Estimation:** We introduce a trustworthiness estimation mechanism that aggregates retrieval confidence, critic confidence, and reflexion burden into a unified reliability score for uncertainty-aware grading.
3. **Evidence-Grounded Multi-Agent Evaluation:** We develop a retrieval-grounded Actor–Critic Reflexion framework where grading decisions are explicitly constrained by retrieved course evidence and independently verified for factual consistency and rubric alignment.
4. **Selective Automation Analysis for Educational Assessment:** We empirically evaluate the relationship between trustworthiness estimation, hallucination reduction, and grading reliability under selective automation settings, demonstrating effective identification of uncertain evaluations requiring human oversight.

The remainder of this paper is organized as follows: Section 2 reviews related work in LLM-based assessment and RAG systems. Section 3 details the proposed system architecture, including the extraction, retrieval, and multi-agent grading components. Section 4 presents our experimental methodology and evaluation metrics and discusses the results and performance of the system, and Section 5 concludes the paper with directions for future research.

2 RELATED WORK

Automated educational assessment has evolved from rule-based grading systems to modern AI-driven evaluation frameworks. This section reviews prior work in traditional automated grading, LLM-based assessment, retrieval grounding, multi-agent reasoning, and selective automation.

2.1 Classical Machine Learning and NLP-Based Grading

The first automated grading tools used the keyword approach, which was not efficient enough since it did not take into account the semantics. To overcome these difficulties, NLP-based grading solutions used such techniques as synonym expansion, contextual matching, and ontological representation (Rokade et al., 2018). Geigle et al. (Geigle et al., 2016) have shown how active learning may provide reliable grading with fewer human interventions.

Further work used more advanced methods such as Naïve Bayes, Support Vector Machines, and Random Forests that showed better results compared to the rule-based methods (Xu, 2024). Larger reviews were conducted by Kumar et al. (Kumar et al., 2024) and Wang et al. (Wang et al., 2025), and they point out the increasing importance of educational data mining and performance evaluation.

2.2 Large Language Models in Educational Assessment

On the other hand, the development of LLMs such as GPT-3(Brown et al., 2020) and GPT-4(OpenAI, 2023) has achieved great progress in improving automated assessment. These LLMs have demonstrated outstanding abilities to reason through few-shot and zero-shot learning techniques and thus became promising tools for automatic assessments.

Xu *et al.*(Xu et al., 2024) present an in-depth investigation of the use of LLMs in education with a focus on tutoring, feedback delivery, and scalable assessment.

There is evidence that suggests the applicability of LLMs to automated grading processes. For instance, Uyar and Büyükahıska (Uyar and Büyükahıska, 2025) evaluated the effectiveness of ChatGPT as an AES by correlating the output of the system with human evaluators based on widely recognized rubrics (such as IELTS). Nonetheless, Wang *et al.*(Wang et al., 2025, 2024) advise against using such systems because of potential issues related to biases, lack of explainability, and inaccurate feedback.

2.3 Grounding Evaluations with Retrieval-Augmented Generation (RAG)

RAG increases factual grounding through integration of parametric LLM knowledge with retrieved facts (Lewis et al., 2020). According to Lewis *et al.* (Lewis et al., 2020), the idea behind RAG is to build a hybrid model incorporating parametric knowledge integrated with LLM and non-parametric one that is retrieved using dense vector indices. In this way, LLM can retrieve external sources to support the answers it generates.

Generation models powered by relevant passage retrieval demonstrate a notable increase in factual accuracy in QA (Izacard and Grave, 2020). As for educational evaluation, RAG helps anchor evaluation judgments to validated citations on the basis of teaching material and rubrics. On the other hand, Gao et al. (2023) suggest post-processing the LLM output with evidence retrieval to correct its mistakes, but unlike them, we ground evaluation exclusively in a domain-specific database.

2.4 Advanced Prompting and Multi-Agent Reflexion

While RAG enhances grounding, complex problems need reasoning. CoT prompting was proposed by Wei *et al.* (Wei et al., 2022), which showed that explicitly reasoning through intermediary steps greatly enhanced the ability to solve logic and arithmetic questions.

The reasoning-then-acting paradigm (Yao et al., 2023b) underpins our Actor agent, while multi-agent architectures such as AutoGen (Wu et al., 2023) facilitate agent cooperation and Reflexion (Shinn et al., 2023) enables iterative self-correction without retraining. Tree-structured reasoning (Yao et al., 2023a) offers an alternative deliberation strategy, though our design opts for a linear Actor-Critic loop for computational efficiency.

2.5 Human-in-the-Loop and Selective Automation

Studies in human-AI interaction highlight how deferring difficult predictions to human experts may result in increased reliability of high-stakes decisions (Geifman and El-Yaniv, 2017). Selective prediction models demonstrate that avoiding predictions of low confidence may lead to higher reliability of automated systems (Geifman and El-Yaniv, 2017).

On the other hand, current automated scoring systems using LLMs, retrieval-based augmentation, and multi-agent reasoning methods always produce unconditional grades without taking grading uncertainty into account and assessing whether human verification is needed (Uyar and Büyükahıska, 2025; Lewis et al., 2020; Shinn et al., 2023). Thus, there is a potential limitation to their applicability since wrong grades might negatively impact the process of evaluation.

In order to mitigate such problems, we propose a multi-signal trustworthiness estimator based on the confidence of retrieval and critic agents and the reflexion burden.

2.6 Summary and Research Gap

Automated grading methods have evolved from NLP-based semantic similarity approaches (Ayaan and Ng, 2025) to large language model systems (Uyar and Büyükahıska, 2025). More advanced techniques, including retrieval-augmented generation (Lewis et al., 2020) and reflexion-based multi-agent frameworks (Shinn et al., 2023), improve semantic understanding, factual grounding, and iterative correction. However, existing systems generally focus on grading accuracy and assume all generated evaluations can be accepted automatically, without explicitly modeling grading uncertainty.

Consequently, a key research gap remains in trust-aware assessment: determining whether a generated evaluation is sufficiently reliable for autonomous grading or should be referred for human review. This work addresses that gap through a selective grading framework that integrates retrieval grounding, Actor–Critic Reflexion, and trustworthiness estimation to identify confident and uncertain evaluations.

3 Methodology

3.1 System Overview

In this research, we propose a retrieval-based LLM-based system for automatic assessment that compares students' answers with respect to a trustworthy knowledge base of the course (Rokade et al., 2018; Kumar et al., 2024; Wang et al., 2025; Xu et al., 2024; Lewis et al., 2020). It is based on four modules including hybrid document parsing, two-step dense retrieval with cross-encoder re-ranking (Izacard and Grave, 2020), evaluation with Actor–Critic multi-agent system (Wu et al., 2023; Shinn et al., 2023)

3.2 Hybrid Heuristic-Based Document Parsing

The student submissions may include a combination of digitized pages as well as scanned pages, making automatic text extraction difficult (Xu, 2024; Ayaan and Ng, 2025). In order to achieve maximum text extraction precision, a hybrid parsing method based on heuristics is applied.

Starting with the input document D comprising of n pages, native text extraction from each page p_i is carried out using *PyMuPDF*. The resulting extracted text will be represented as $T(p_i)$.

In order to identify the scanned pages or pages with bad encoding, a heuristic approach based on the text length is applied. If $|T(p_i)| < \tau$, then the page is considered to be a likely candidate for scanning and OCR using Tesseract through the *pytesseract* library is automatically activated (Xu, 2024; Ayaan and Ng, 2025). This is done in order to restore the lost information in the text extraction process. The extraction algorithm can be formally described as

$$E(p_i) = \begin{cases} T(p_i), & \text{if } |T(p_i)| \geq \tau, \\ \text{OCR}(p_i), & \text{if } |T(p_i)| < \tau. \end{cases}$$

3.3 Knowledge Base Construction

Knowledge base construction is done offline using the reliable course material such as textbooks, video lectures, and reference documents, which constitute the knowledge base used in the retrieval pipeline (Kumar et al., 2024; Wang et al., 2025; Xu et al., 2024). The source document is segmented into segments using the sliding window technique to maintain semantic continuity among segments (Lewis et al., 2020; Izacard and Grave, 2020). The source document S can be considered as a string of m characters: $S = (c_1, c_2, \ldots, c_m)$ We define chunk length $W = 1000$ and overlap $O = 200$ characters, giving stride $\Delta = W - O = 800$. The i-th chunk is:

$$C_i = (c_{i\cdot\Delta}, \ldots, c_{i\cdot\Delta+W-1}), \quad i \in \{0, 1, 2, \ldots\}$$

Each chunk is then encoded into a dense vector representation using `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019), mapping each text segment to a d-dimensional semantic space:

$$f_{\text{emb}} : \mathcal{X} \to \mathbb{R}^d, \qquad v_i = f_{\text{emb}}(C_i)$$

The resulting vectors, along with their corresponding text content and metadata, are stored in a local indexed repository. Formally, the knowledge base is defined as the set of all chunk-vector pairs:

$$\mathcal{KB} = \{(C_i, v_i, \mathcal{M}_i) \mid 0 \leq i \cdot \Delta \leq m - W\}$$

where $\mathcal{M}_i$ denotes the metadata associated with chunk C_i, such as source document name and page number.

3.4 Two-Stage Dense Retrieval with Cross-Encoder Re-Ranking

The relevance-based context is created via a two-step pipeline that involves dense retrieval to retrieve candidate entities and then semantic re-ranking (Lewis et al., 2020; Izacard and Grave, 2020). At the initial step, the input query q and knowledge chunk embeddings are produced by the shared embedding model f_{emb} (Vaswani et al., 2017; Brown et al., 2020; OpenAI, 2023):

Cosine similarity is then computed between the query vector v_q and all stored chunk embeddings $v_i \in \mathcal{KB}$ to produce an initial relevance ranking. The bi-encoder similarity score s_{bi} is formally defined as:

$$s_{\text{bi}}(q, C_i) = \frac{v_q \cdot v_i}{\|v_q\| \|v_i\|}$$

The top N candidates ($N = 20$) are selected:

$$\mathcal{C}^{(1)} = \underset{C_i \in \mathcal{KB}}{\arg\max}^{(N)} s_{\text{bi}}(q, C_i)$$

To capture richer query-document interactions, the retrieved candidates are re-ranked using a cross-encoder model that jointly processes q and each candidate passage $C \in \mathcal{C}^{(1)}$ (Izacard and Grave, 2020).

The refined score is computed as:

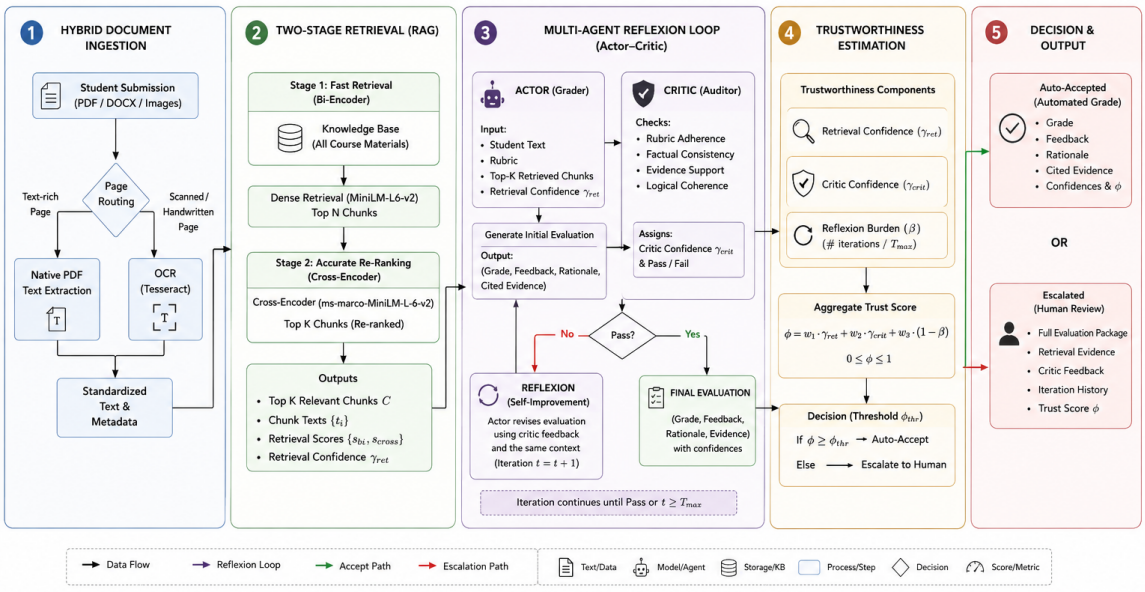

Figure 1: Overview of the proposed full pipeline and methodology for fact-grounded automated grading.

$$s_{\text{cross}}(q, C) = f_{\text{cross}}(q, C)$$

The passages are then re-ordered according to these refined scores, and the top K segments (default value $K = 5$) are selected as the final Truth Context $\mathcal{C}^{(2)}$ used during grading:

$$\mathcal{C}^{(2)} = \underset{C \in \mathcal{C}^{(1)}}{\arg\max}^{(K)} s_{\text{cross}}(q, C)$$

To determine the best trade-off between semantic recall and computational efficiency without the context being overly large for the LLM, the default retrieval parameters were chosen based on the assumption that N=20 and K=5.

The confidence of the extracted evidence is estimated by the retrieval module using a retrieval confidence score:

$$\gamma = \frac{1}{K} \sum_{i=1}^{K} s_{\text{cross}}\left(q, C_i^{(2)}\right) \tag{1}$$

where $\gamma \in [0, 1]$ denotes retrieval confidence. Higher values indicate stronger semantic alignment between the query and retrieved evidence. This score is later used by the trustworthiness estimation module.

3.5 Trustworthiness Estimation

In the traditional method of automation, the uncertainty level is never taken into consideration. The traditional method always results in the generation of an assessment. To achieve selective automation, we present a trustworthiness estimation module that will decide, based on different information sources, whether the generated assessment is trustworthy or not.

The trust model combines three signals: retrieval confidence γ (Section 3.4), critic confidence c, and a refinement burden score based on the proportion of Reflexion iterations used:

$$\beta = \frac{\text{Reflexion Iterations}}{\text{Maximum Allowed Iterations}} \tag{2}$$

where $\beta \in [0, 1]$, with larger values indicating greater evaluation difficulty.

These signals are aggregated into a unified trustworthiness score:

$$\delta = w_1\gamma + w_2 c + w_3(1 - \beta) \tag{3}$$

subject to $w_1 + w_2 + w_3 = 1$,

where $\delta \in [0, 1]$ represents the overall reliability of the generated evaluation.

A threshold-based decision rule is then applied:

$$\text{Output} = \begin{cases} E_{\text{final}}, & \delta \geq \phi, \\ \text{FLAG}, & \delta < \phi, \end{cases} \tag{4}$$

where ϕ is a predefined trust threshold. Evaluations exceeding the threshold are accepted automatically, while lower-confidence cases are flagged for instructor review.

3.6 Multi-Agent Evaluation Architecture

To improve grading reliability, we employ a multi-agent LLM architecture that separates generation and verification, reducing hallucinations and ensuring evidence-grounded evaluation (Wu et al., 2023; Shinn et al., 2023).

The process begins with an **Actor** agent, which performs the primary grading task. Let the student response be A, the grading rubric be R, and the retrieved context be $\mathcal{C}^{(2)}$. The Actor, modeled as $\mathcal{M}_{\text{Actor}}$, produces an initial evaluation:

$$E_0 = \langle m, r, e \rangle = \mathcal{M}_{\text{Actor}}(A, R, \mathcal{C}^{(2)})$$

where m is the score, r the reasoning, and e the supporting evidence.

A **Critic** agent then independently audits this output for rubric alignment, factual consistency, and logical correctness (Shinn et al., 2023). In addition to a textual critique t_c, it outputs a validity indicator $v \in \{0, 1\}$ and confidence score $c \in [0, 1]$:

$$O_C = \langle v, c, t_c \rangle = \mathcal{M}_{\text{Critic}}(E_0, A, R, \mathcal{C}^{(2)})$$

This separation of roles enables systematic detection of errors and supports more consistent and grounded grading. The iterative critique and revision process is described in the next section.

3.7 Reflexion-Based Critique Loop

Single-pass automated grading may produce incomplete or poorly justified evaluations (Wei et al., 2022). To improve reliability, the framework incorporates a Reflexion-based critique loop that enables iterative self-correction (Shinn et al., 2023).

Following the generation of an initial assessment E_0 by the Actor and audit $O_C = \langle v, c, t_c \rangle$ by the Critic, the system checks whether revision is needed (Wu et al., 2023; Shinn et al., 2023). A value $\theta = 0.85$ serves as the threshold for evaluating the tradeoff between grade accuracy and computational efficiency. In cases when the result of evaluation is invalid ($v = 0$) or the confidence level does not satisfy the constraint ($c < \theta$), the process invokes a Senior Grader agent for revision using $(A, R, \mathcal{C}^{(2)})$.

The revised evaluation is generated as

$$E_{\text{revised}} = \mathcal{M}_{\text{Senior}}(A, R, \mathcal{C}^{(2)}, E_0, O_C)$$

The final output of the Reflexion loop is

$$E_{\text{final}} = \begin{cases} E_0, & \text{if } v = 1 \text{ and } c \geq \theta \\ E_{\text{revised}}, & \text{otherwise.} \end{cases}$$

This mechanism serves as a quality-control layer that improves rubric alignment and evaluation reliability.

3.8 Evidence-Grounded Scoring

The grading framework adopts an evidence-grounded scoring strategy in which all evaluation decisions are supported by retrieved context (Lewis et al., 2020; Izacard and Grave, 2020).

Given a rubric $R = \{r_1, r_2, \ldots, r_k\}$, where each criterion r_j has maximum weight W_j, the agent evaluates the student response A against the retrieved context $\mathcal{C}^{(2)}$. For each criterion, an alignment function estimates the degree to which the response satisfies the rubric requirements (Liu et al., 2023):

$$f_{\text{align}}(A, r_j, \mathcal{C}^{(2)}) \to \alpha_j$$

where $\alpha_j \in [0, 1]$ denotes factual correctness and completeness. The awarded marks are computed as

$$m_j = \alpha_j W_j$$

For interpretability, each criterion produces an explanation tuple $T_j = \langle m_j, \rho_j, \epsilon_j \rangle$, where ρ_j is a justification trace for the assigned score (Wei et al., 2022), and $\epsilon_j \subseteq \mathcal{C}^{(2)}$ is an evidence quote supporting the evaluation.

The final score is

$$M = \sum_{j=1}^{k} m_j$$

yielding transparent and evidence-backed grading.

3.9 Structured Output Schema

The system outputs all grading results in a structured format to support transparency and downstream analysis.

Let $Q = \{q_1, q_2, \ldots, q_n\}$ denote the set of evaluated questions. Each question q_i is represented by an evaluation entry $\mathcal{G}_i$ containing the assigned marks m_i, an aggregated reasoning trace ρ_i, and supporting evidence ϵ_i (Lewis et al., 2020; Wei et al., 2022):

$$\mathcal{G}_i = \langle m_i, \rho_i, \epsilon_i \rangle, \qquad \mathcal{G} = [\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n]$$

This structure provides traceable and evidence-supported grading results (Xu et al., 2024; Uyar and Büyükahıska, 2025).

The metadata component $\mathcal{M}$ stores processing statistics, parsing information, and extraction details. The complete output object is defined as

$$O = \langle \mathcal{G}, F, S, \delta, \mathcal{M} \rangle$$

where $\mathcal{G}$ contains grading results, F generated feedback, S supporting evidence, δ the trustworthiness score, and $\mathcal{M}$ system metadata.

The overall grading function is defined as

$$\Phi : \mathcal{D} \times \mathcal{R} \rightarrow (O, \delta)$$

3.10 System Metrics and Logging

For the document parsing stage, the system records whether each page is processed using native text extraction or OCR. The OCR utilization rate μ_{OCR} is defined as:

$$\mu_{\text{OCR}} = \frac{1}{n} \sum_{i=1}^{n} I(|T(p_i)| < \tau) \tag{5}$$

where $T(p_i)$ denotes the extracted text from page p_i and τ is the minimum text threshold triggering OCR fallback.

The evaluation module records audit outcomes v, Critic confidence scores c. For a set of evaluated questions Q, the Reflexion trigger rate is:

$$\mu_{\text{Reflexion}} = \frac{1}{|Q|} \sum_{j=1}^{|Q|} I(v_j = 0 \vee c_j < \theta) \tag{6}$$

where θ denotes the confidence threshold used to trigger revision.

To quantify hallucination reduction, we define the Evidence Hallucination Rate:

$$H = \frac{1}{|Q|} \sum_{j=1}^{|Q|} I(v_j = 0) \tag{7}$$

where lower values indicate stronger factual grounding and closer adherence to the retrieved Truth Context.

To characterize selective automation, the framework records the flagging rate:

$$\mu_{\text{flag}} = \frac{1}{|Q|} \sum_{j=1}^{|Q|} I(\delta_j < \phi) \tag{8}$$

where δ_j is the trustworthiness score associated with evaluation j and ϕ is the trust threshold. The complementary automation coverage rate is: $\mu_{\text{auto}} = 1 - \mu_{\text{flag}}$ Finally, total processing latency is measured as:

$$T_{\text{total}} = T_{\text{ingest}} + T_{\text{retrieve}} + T_{\text{evaluate}} \tag{9}$$

All recorded system metrics are aggregated into the metadata object:

$$M = \langle \mu_{\text{OCR}}, \mu_{\text{Reflexion}}, H, \mu_{\text{flag}}, \mu_{\text{auto}}, T_{\text{total}}, \ldots \rangle. \tag{10}$$

4 Results

This algorithm was evaluated using a dataset of 1000 assignments of students from varied subjects like Distributed Computing File Systems, Humanities, and Marketing. This dataset is inclusive of both technical and non-technical assignments, which evaluates the generalizing ability of this algorithm on different criteria that are considered while grading the assignments. Each assignment is marked based on a set rubric having various criteria and fixed weight distributions defined by the instructor. All these assignments have been marked based on this rubric only, which ensures consistency across the assignments of each subject.

Let y_i be the true mark for a particular sample, while $\hat{y}_i$ be the predicted mark from the system. There are $N = 1000$ samples.

4.1 Evaluation and Results

In terms of evaluating the accuracy of the grades, we use metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and others for regression, as well as correlation coefficient. The accuracy of our approach gives 1.9 MAE, 2.31 RMSE, and 0.987 correlation coefficient.

4.2 Trustworthiness-Aware Selective Grading

In addition to accuracy, another criterion that is considered is the ability of the proposed model to identify any evaluation which is not trustworthy enough and needs human verification. This will be based on a parameter called trustworthiness δ in relation to a threshold value ϕ.

Table 1 shows the tradeoff between automation coverage and grading reliability under different threshold settings.

Table 1: Selective grading performance under varying trustworthiness thresholds.

Threshold ϕ	Auto-Grade (%)	Flagged (%)	MAE (Auto)	MAE (Flagged)
0.70	88	12	1.5	5.2
0.80	76	24	1.2	4.9
0.90	61	39	0.9	4.7

The trustworthy estimator proposed is quite effective in distinguishing between dependable assessments and challenging assignments. The assignments marked by the system for human grading always have much higher levels of grading error, meaning that the system is indeed able to identify evaluations that are not sure enough and cannot be automatically graded. This selective grading approach is one of the main contributions of the whole system framework.

4.3 Comparison with Zero-Shot Baselines

To assess the effectiveness of the presented framework, we conduct comparative analysis by comparing the performance of the framework against popular LLMs including GPT-5, Claude Opus 4.6, and Gemini 3 for the same dataset that consists of 1000 student responses along with the corresponding grading rubrics. In this case, the student response and grading rubric are directly fed to the LLM without employing retrieval grounding and multi-agent validation techniques.

On the other hand, our approach extends rubric assessment with retrieval grounding (Lewis et al., 2020) and a systematic reflection framework for multiple agents (Shinn et al., 2023), thereby facilitating direct validation of evidence and rectification of any errors. Despite obtaining reasonable correlations among scores, zero-shot LLMs using rubrics display high variance and errors, signifying susceptibility to different interpretations of prompts and absence of facts (Gardner et al., 2021).

4.4 Qualitative Analysis.

We have noticed that zero-shot LLMs tend to provide unsourced or paraphrased evidence during the evaluation of open-ended answers. Our proposed solution, on the other hand, provides evidence-based rationale derived from retrieved information, thus minimizing inaccuracies in facts. But sometimes, there may be errors when the retrieval process fails or OCR extraction is inadequate.

Table 2: Quantitative Comparison with Zero-Shot Baselines

Model	MAE	RMSE	Var	Corr	Hallucination Rate (%)
Proposed Framework	**1.9**	**2.31**	**207.2**	**0.987**	**2.9**
GPT-5 (Zero-shot)	3.43	4.30	223.51	0.9576	12.3
Claude Opus 4.6 (Zero-shot)	3.34	4.16	224.04	0.96	7.1
Gemini 3 (Zero-shot)	3.24	4.02	216.55	0.9619	9.9

4.5 Ablation Study

To understand the contribution of each component in the proposed framework, we perform an ablation study by systematically removing key modules.

Table 3: Ablation study evaluating the impact of individual components.

Model Variant	MAE	RMSE	Hallucination Rate (%)
Full Framework	**1.90**	**2.31**	**2.9**
w/o RAG	3.04	3.93	11.6
w/o Critic	2.66	3.35	8.7
w/o Reflexion	2.28	2.89	6.4
LLM only	3.42	4.27	13.1

5 Conclusion

The solution suggested by us is a robust autograder using two-stage RAG pipelines (Lewis et al., 2020) and a reflective multi-agent model (Shinn et al., 2023). The hybrid parser, evaluation through retrieval, and verification through actor-critic help avoid potential problems associated with zero-shot LLMs (Xu et al., 2024), including hallucinations and rubric compliance.

Results from evaluation on 1000 essays assigned in three different domains reveal extremely high similarity with the grades given by the teacher with MAE=1.9, RMSE=2.31, and correlation coefficient=0.987. The most significant achievement of this model is the fact that it ensures that there is no more than 2.9% evidence hallucination.

In light of this, whereas retrieval grounding ensures truthfulness, and reflexion facilitates improvement continuously, future research shall be focused on minimizing the computation delays necessary for MOOCs implementation in real time, and expanding the method to evaluate multimedia student input.

The suggested framework based on trust does not treat grading results equally. The proposed grading solution is capable of differentiating confidently automatable cases from the uncertain ones, thus, making its use more realistic in educational settings by applying this selective automation approach.

Limitations

While the proposed Two-Stage RAG and Multi-Agent Reflexion framework improves grading reliability and factual grounding, several limitations remain.

Computational Overhead: The multi-phase approach increases computational costs through the retrieval process, re-ranking process, and the inference stages for Actor, Critic, and potentially refinement of responses. This might affect scalability in live and scaled deployment environments.

Dataset Scale and Generalizability: This model has been tested using a data set of 1000 graded assignments from various fields. However, even though this model demonstrates excellent consistency with the grades provided by the instructors, it needs to be validated using more extensive data sets before being considered universally applicable.

Handling of Visual Content: In the current system, text extraction through native parsing and OCR is employed. The method is efficient for textual input, but not so effective when dealing with diagrams and flowcharts. In some situations, OCR will generate inaccurate output, thus preventing full comprehension of the student's response.

Dependence on Retrieval Quality: Performance in grading is based on the nature and scope of knowledge that exists. If the relevant reference sources are not available, then this may impact negatively on the result obtained from context retrieval.

The current trustworthiness estimator relies on empirically selected aggregation weights and threshold values, which may require calibration across institutions and grading settings.

Fixed Hyperparameters: The system uses predefined thresholds (e.g., OCR fallback and Reflexion confidence), which may require tuning for different datasets or document qualities.

References

A. Ayaan and K.-W. Ng. 2025. Automated grading using natural language processing and semantic analysis. *MethodsX*, 14:103395.

T. B. Brown and 1 others. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y Zhao, Ni Lao, Hung-yi Lee, Da-Cheng Juan, and 1 others. 2023. RARR: Researching and revising what language models say using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 16856–16893.

J. Gardner, M. O'Leary, and L. Yuan. 2021. Artificial intelligence in educational assessment: 'breakthrough? or buncombe and ballyhoo?,'. *Journal of Computer Assisted Learning*, 37(5):1207–1216.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

C. Geigle, C. Zhai, and D. Ferguson. 2016. An exploration of automated grading of complex assignments. In *Third (2016) ACM Conference on Learning @ Scale (L@S '16)*.

G. Izacard and E. Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

C. V. A. Kumar, A. K. Eemani, G. C. Kalluri, and G. Rudra. 2024. A survey on automated student evaluation and analysis using machine learning. *World Journal of Advanced Research and Reviews*, 21(2):2547–2554.

P. Lewis and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3214–3252.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

A. Rokade, B. Patil, S. Rajani, S. Revandkar, and R. Shedge. 2018. Automated grading system using natural language processing. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 1123–1127, Coimbatore, India.

I. Saputra, A. Kurniawan, M. Yanita, E. Y. Putri, and M. Mahniza. 2024. The evolution of educational assessment: How artificial intelligence is shaping the trends and future of learning evaluation. *Indonesian Journal of Computer Science*, 13(6).

N. Shinn, F. Cassano, A. Gopinath, K. R. Narasimhan, and S. Yao. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*.

B. B. Tomić, A. D. Kijevčanin, Z. V. Ševarac, and J. M. Jovanović. 2023. An ai-based approach for grading students' collaboration. *IEEE Transactions on Learning Technologies*, 16(3):292–305.

A. C. Uyar and D. Büyükahıska. 2025. Artificial intelligence as an automated essay scoring tool: A focus on chatgpt. *International Journal of Assessment Tools in Education*, 12(1):20–32.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, and 1 others. 2017. Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*.

J. Wang, Y. Hussain, X. Zhang, C. Mao, and Z. Wu. 2025. Ai-based automated grading systems: Opportunities, challenges, and future directions. *Business and Social Sciences Proceedings Review (GEEA)*.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

J. Wei and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Q. Wu and 1 others. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

D. C. F. Xu. 2024. Design and implementation of machine learning algorithms in automatic grading of students' assignments. *Journal of Electrical Systems*, 20-3s:899–919.

H. Xu, W. Gan, Z. Qi, J. Wu, and P. S. Yu. 2024. Large language models for education: A survey. *arXiv preprint arXiv:2405.13001*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.