Human-in-the-Loop Short-Answer Grading with Calibrated Confidence and Adaptive Rubric Alignment

Anonymous ACL submission

Abstract

Recent advances in large language models (LLMs) have enabled promising progress in automatic short answer grading (ASAG). However, most approaches overlook the fact that grading rubrics and styles evolve, leading to unreliable and outdated evaluations. Furthermore, ASAG systems based on LLMs can produce overconfident evaluations, highlighting the need for post-hoc calibration to ensure fair and trustworthy assessment. We propose a Human-in-the-Loop (HiL) ASAG system that integrates calibrated confidence estimation with selective human oversight. The model abstains from low-confidence cases, which are then reviewed by the teachers. We find that human corrections not only enhance grading reliability and provide supervision for continuous fine-tuning, enabling the model to adapt to new rubrics and grading conventions. Confidence calibration through temperature scaling ($T \approx 1.8$) reduces the Expected Calibration Error from 0.215 to 0.055, resulting in smoother confidence distributions and more reliable acceptance behavior. Moreover, a single HiL iteration further improves reliability and coverage, focusing human effort on true errors while maintaining grading quality. Overall, our approach combines scalability with human judgment, enabling trustworthy and adaptive assessment in educational contexts.

1 Introduction

Assessment plays a crucial role in the learning process, but it can become demanding when scaled, particularly with formative assessments designed to guide both students and teachers. Traditional closed-response formats, such as multiple-choice questions, are easy to grade but often fail to capture genuine understanding. Open-ended questions, on the other hand, provide deeper insight, yet grading them is time-consuming and impractical at scale. Automatic Short Answer Grading (ASAG) addresses this challenge by offering fast and consistent evaluations, making it a promising approach for improving assessments without sacrificing quality (Ferreira Mello et al., 2025).

Early approaches using Machine Learning (ML) typically framed the task as a supervised regression or classification problem, using TF-IDF vectors or sentence embeddings, and training models, such as Random Forests, Support Vector Machines, and various regressors to predict student performance scores (Weegar and Idestam-Almquist, 2024; del Gobbo et al., 2023). With the advent of Deep Learning (DL), neural network models such as CNNs, LSTMs, and BiLSTMs, often combined with embeddings like BERT or Skip-Thought, enabled rich semantic representations of student responses and achieved higher alignment with teacher grading (Mardini et al., 2024; Prabhudesai and Duong, 2019; Abdul Salam et al., 2022; Mirjalili et al., 2014; Ljungman et al., 2021). These approaches demonstrated that ML models could achieve substantial agreement with human teachers.

Nonetheless, current ML and DL approaches for ASAG suffer from two main limitations: **(i)** they lack adaptability, as they offer no mechanisms to incorporate evolving teacher judgments or rubric changes; and **(ii)** as a result, they require costly retraining to adapt to new content or grading policies.

To the best of our knowledge, the question of how to build ASAG systems that can adapt dynamically, minimize human effort, and produce calibrated, trustworthy grade predictions remains unresolved.

Motivated by this gap, the main contributions of this paper are as follows:

- **We propose a Human-in-the-Loop (HiL) ASAG system that combines confidence calibration with targeted human oversight for adaptive automatic grading.** The system consists of two components: a Llama-Instruct

model that performs the grading and a post-hoc calibration component that produces calibrated confidence scores for each grade prediction.

- **We demonstrate that the system achieves human-comparable grading reliability through calibrated confidence estimation and targeted teacher intervention.** It does so by directing teacher attention toward exam responses where human judgment is most beneficial, i.e., those with low confidence or ambiguous content; thus reducing correction effort and maintaining consistent performance even as grading rubrics and question formats evolve across exams.

- **We conduct large-scale simulations of the proposed system on a real-world dataset replicating real exam workflows to evaluate grading reliability and efficiency.** Using short-answer exams from a master's-level Data Mining course (2020–2023), we simulate realistic grading dynamics, including rubric evolution and teacher variability. Our results suggest that the proposed system achieves 70% of the grading task automatically with strong agreement between automated and human grades.

2 Related Work

Recently, Impey et al. (2025) employed a GPT-4 model for grading in Massive Open Online Courses (MOOCs), prompting it with specific rubrics to produce scores that are indistinguishable from those of teachers, outperforming peer grading. Moreover, Aggarwal et al. (2025) augmented plain grading with elaborated feedback by fine-tuning LLaMA-2 (13B) and Mistral-7B, addressing the limitation of most ASAG systems, which fail to provide content-focused feedback. For Automating Essay Scoring (AES), which consists of assessing longer student responses, Zeinalipour et al. (2025) proposed a method to provide fine-grained automated syntax feedback to improve students' writing proficiency, beyond simple scoring. This was achieved by prompting GPT3.5-Turbo to generate syntax feedback from the collected data, and then fine-tuning LLMs to enhance their ability to provide syntax feedback.

Furthermore, several recent studies have incorporated Retrieval-Augmented Generation (RAG) to inject external knowledge into the LLMs, thereby improving factual accuracy, expanding their knowledge base, and enhancing system flexibility for grading and feedback on new content. More specifically, Duong and Meng (2024) used lecture notes that are fragmented, embedded, and retrieved to guide the grading process performed by a GPT-3.5/4 model. Using few-shot prompting with low/medium/high quality examples, the model improved Pearson correlations with human evaluators compared to an embedding-based solution and state-of-the-art ASAG systems. In another example of RAG integration for ASAG with LLMs, Chu et al. (2025) constructed a triple-index knowledge base including knowledge ranging from general science documents to expert-annotated graded examples with feedback. Given a task, a question, a rubric, retrieved knowledge, and examples to be graded, the LLM grader achieved consistent improvements in F_1 scores and accuracy in zero-shot settings compared to the Non-RAG system.

The systems mentioned above still lack the ability to adapt to new rubrics and grading policies, as well as the inclusion of the teacher in the grading process. In an attempt to address these limitations, Gorbatovski and Kovalchuk (2024) utilized Reinforcement Learning from Human Feedback (RLHF) in programming question-answers, leveraging Stack Overflow votes as a scalable source of human feedback to fine-tune GPT-Neo 125M. Their system yielded considerable results when compared to larger models, highlighting the impact of fine-tuning with human-derived feedback. While effective in aligning model outputs with community preferences, this approach relied on indirect and noisy signals and remained static after training. In contrast, Woodrow et al. (2025) optimized LLM feedback to align with direct teacher preference optimization (DPO) in real classroom settings where assistants selected or revised feedback produced by an LLM. However, both approaches retain key limitations: RLHF depends on proxy feedback rather than explicit expert judgment, and DPO requires sustained human annotation effort, and neither incorporates calibration or abstention mechanisms needed for trustworthy, adaptive grading.

3 A Human-in-the-Loop ASAG system

In this paper, our main focus is on short-answer grading, a task that requires evaluating not just factual correctness but also the quality and com-

pleteness of students' responses. Short answers offer a rich middle ground between multiple-choice and essay questions, as they allow for open-ended reasoning while remaining structured enough for automated assessment. The overall architecture of the proposed HiL ASAG system is illustrated in Figure 1.

3.1 Model Training

Instruction-tuning Our primary approach to the ASAG task is instruction-tuning, as it enables large language models to follow grading rubrics faithfully and maintain consistent in-domain performance (Zhang et al., 2025). Compared to alternative strategies, instruction-tuning provides a more robust and interpretable alignment between model behavior and human grading criteria. Other approaches fall short in important ways. Treating grading as *prompted generation* (Brown et al., 2020, in zero-shot or few-shot configurations) often works well for general NLP tasks such as question answering, summarization, or translation, but yields inconsistent and unreliable results in real-world grading scenarios. Similarly, *classification-based grading* may appear straightforward but struggles with question-specific grading scales (different *MaximumGrade* ranges), where normalization tends to obscure the intended meaning of the rubric.

Teacher Forcing Our instruction-tuning approach constrains the model to emit structured responses of the form `{"grade": g, "max_grade": M}` where $g \in \{0, 1, \ldots, M\}$ and M is the question-specific maximum grade. This format ensures consistent parsing across heterogeneous question scales while maintaining the ordinal semantics of grades. The training data consists of question–answer pairs with human-provided grades, formatted using standardized prompts that explicitly request outputs following the grading rubric.

At inference, we score all candidate grades deterministically using teacher forcing to avoid sampling variability. For each grade $g \in \{0, \ldots, M\}$, we compute the model's likelihood of assigning that grade by conditioning on the corresponding structured response and summing the token-level log probabilities:

$$z_g = \sum_{t=1}^{L_g} \log p_\theta\big(y_{g,t} \mid x, y_{g,<t}\big) \tag{1}$$

where $y_{g,t}$ is the t-th token in the response for grade g, L_g is the length of that response, and x is the input question–answer pair. This produces a logit vector $\mathbf{z}$ where each element represents the model's confidence for the corresponding grade.

Temperature Scaling We apply temperature scaling (Xie et al., 2024) to convert raw logits to calibrated class probabilities, from which we derive the prediction and confidence:

$$p_g(T) = \frac{\exp(z_g/T)}{\sum_{h \in G} \exp(z_h/T)} \tag{2}$$

$$\hat{g} = \arg\max_{g \in \{0..M\}} p_g(T) \tag{3}$$

$$\hat{C} = \max_{g \in \{0..M\}} p_g(T) \tag{4}$$

where T is the temperature parameter, $\hat{g}$ is the predicted grade, and $\hat{C}$ is the confidence score used for abstention decisions. Although instruction-tuned models achieve high accuracy, they tend to be over-confident, which motivates the use of post-hoc calibration and a calibrated abstention policy to safely route uncertain cases to human teachers.

3.2 Model Calibration

The Overconfidence Problem After instruction-tuning, models are heavily over-confident: they assign a high confidence score (≈ 0.99) to predictions that are often incorrect. This makes it difficult to trust the model's confidence for routing decisions. We address this through post-hoc temperature scaling, which adjusts the confidence scores to better reflect the model's actual accuracy.

Measuring Calibration Quality We quantify miscalibration using the Expected Calibration Error (Posocco and Bonnefoy, 2021, ECE), which measures how well the predicted confidence matches empirical accuracy. We divide predictions into confidence bins and compute,

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{n_b}{n} |\mathrm{acc}(b) - \mathrm{conf}(b)| \tag{5}$$

where n_b is the number of samples in the bin b, $\mathrm{acc}(b)$ is the empirical accuracy in the bin b, and $\mathrm{conf}(b)$ is the average confidence in bin b. Lower ECE indicates better calibration, with ECE = 0 representing perfect calibration.

We perform a temperature search on a held-out calibration split, with the candidate vector $T \in$

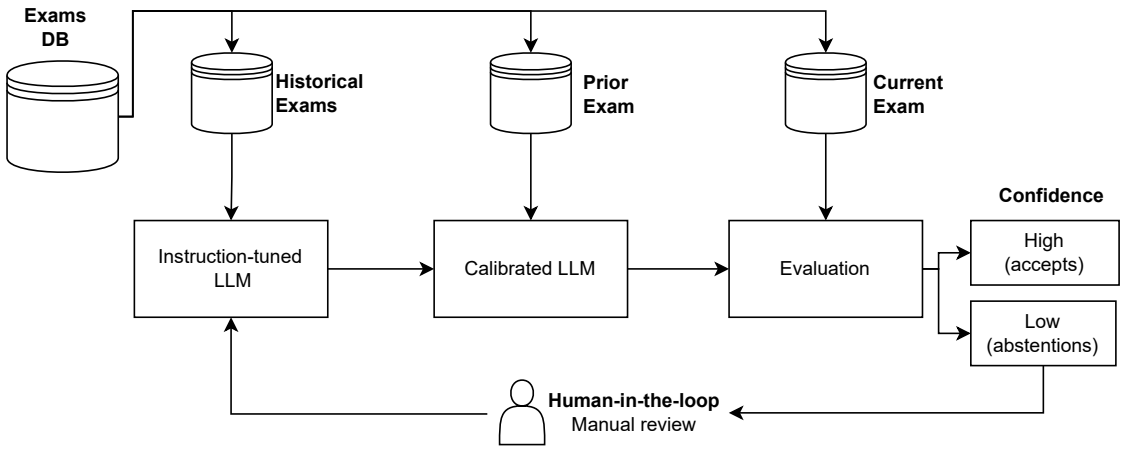

Figure 1: HiL grading system. Historical exams train an instruction-tuning LLM. Prior-year exams calibrate temperature T and select threshold τ. On the current exam, low-confidence cases are abstained to a human reviewer. Corrections, plus a replay buffer from training, drive a conservative model update, followed by recalibration and evaluation.

$[0.1, 3.0]$ to minimize 15-bin ECE. The procedure **(i)** collects logits and ground-truth labels from the calibration split, **(ii)** computes temperature-scaled probabilities for each candidate T, **(iii)** calculates ECE, and **(iv)** selects the T that minimizes ECE. Temperature scaling preserves the model's accuracy while making confidence scores more trustworthy.

Abstention Using reliability diagrams (before/after) and the ECE improvements of the search, we derive the coverage-risk curves (see Figure 3 in the Appendix B) and choose an abstention threshold τ that attains a target coverage. In practice, a target coverage of 70% means: select τ on a calibration split so that approximately 70% of samples (highest calibrated confidences) are auto-accepted and the rest are routed to teachers for correction and HiL fine-tuning. After each HiL iteration, we recalibrate the updated model on the same calibration split to maintain calibration consistency and ensure that subsequent routing and evaluation are based on reliable and trustworthy values.

3.3 Human-in-the-Loop Learning

Low-confidence routed samples are reviewed by teachers, and the resulting corrections form the HiL set H reflecting the current exam's rubric and difficulty. Each correction provides a human-verified grade for a previously abstained sample, creating a targeted training set that addresses the model's specific weaknesses on the current exam. The correction process is designed to be efficient, with teachers focusing only on samples where the model's confidence was low.

To preserve prior performance and stabilize adaptation across heterogeneous scoring scales, we augment H with a replay buffer consisting of stratified samples from historical data matched by *MaximumGrade*. The replay buffer is constructed by sampling an equal number of examples from each *MaximumGrade* category present in the corrections, ensuring that the training distribution matches the correction distribution. This scale-aware sampling prevents the model from overfitting to the most common grade scales and maintains high performance.

Fine-Tuning We fine-tune only the adapters from a Low-Rank Adaptation process (Hu et al., 2022, LoRA) on H with replay while keeping the prompt design fixed to adapt the model to unseen question-answer patterns and recent grading style while controlling overfitting and catastrophic forgetting (Kotha et al., 2024; Luo et al., 2025) through replay. The replay buffer serves as a regularization mechanism. By including historical samples that span various question types, grading styles, and difficulty levels, we ensure that the model's generalization and robustness are optimal.

3.4 Post-HiL Recalibration

After fine-tuning, we recalibrate the updated model on the calibrated split by re-running temperature search to obtain a new optimal temperature T'. This is crucial because the model's confidence dis-

tribution might have shifted during fine-tuning with H being several orders of magnitude lesser when compared to the full training set, potentially rendering the pre-HiL T weak. The re-calibration process follows the same procedure as the intial calibration.

Selecting Thresholds We select the abstention threshold τ, from the corresponding coverage-risk curve to meet the target coverage, apply it for the current phase's evaluation, and re-select it after HiL when recalibration changes the confidence distribution. This avoids leakage (selection never on the test set) while keeping routing aligned with the current model.

We grade the held-out portion of the exam with the re-calibrated model and frozen τ, and report coverage, accepted/abstained counts, accepted set performance, and ECE (pre/post) for each loop. We also analyze the abstentions to quantify adaptation on previously deferred samples. This involves re-evaluating the abstained samples to measure how the model's performance on under-confident and challenging samples has improved after HiL fine-tuning. This provides insights into how effective the correction process was and whether the model was able to learn from the HiL process. If the accepted set performance remains below target, we iterate the HiL: re-select τ, collect new corrections, refresh replay buffer, fine-tune, recalibrate, and re-evaluate until the targets are achieved or the marginal gains have diminished.

4 Experimental Setup

4.1 Data

We use open-ended short-answer based exams from a master's-level Data Mining course administered between 2020 and 2023. The dataset covers a broad range of topics, from machine learning classification and evaluation, clustering, frequent itemset mining to neural networks enabling robust generalization across related topics. The used fields are: *Question*, *Response*, *Grade*, *MaximumGrade*, and *Comment*. Grades are ordinal and question-specific, and we deliberately do not normalize across questions to reflect real-world heterogeneity and evolving rubrics. *MaximumGrade* $M \in \{4, 5, 8, 10\}$ denotes the per-question grading scale (maximum total points): smaller scales (e.g., M=4, 5) typically correspond to concise, assertion-style questions, whereas larger scales (e.g., M=8, 10) are for multi-criterion reasoning questions. The model

Split	Total	M=4	M=5	M=8	M=10
$\mathcal{D}_0$	4000	230	1119	133	2518
$\mathcal{D}_1$	147	0	44	6	97
$\mathcal{D}_2$	378	0	138	12	228

Table 1: Data split sizes and per-M distributions across $\mathcal{D}_0$ (training), $\mathcal{D}_1$ (calibration) and $\mathcal{D}_2$ (evaluation).

produces a structured response that returns the predicted grade g and the corresponding maximum grade M from a standardized prompt template (see Appendix A) so that downstream tasks like scoring, calibration, and abstention operate on a consistent interface. All necessary steps have been taken to ensure that the data does not contain any personally identifying information. More concretely, student names and identifiers have been removed from the text, and the whole corpus has been scrutinized towards preserving student and teacher anonymity.

Chronological Splits To emulate real-world deployment, we partition the data strictly in time: $\mathcal{D}_\mathbf{0}$ **contains historical data (2020-23)** used to train the base model, $\mathcal{D}_\mathbf{1}$ **is the prior year (2024)** split used only for temperature calibration, and $\mathcal{D}_\mathbf{2}$ **is the exam in the current date** held out for final evaluation. We split $\mathcal{D}_2$ 50/50 into $\mathcal{D}_\mathbf{21}$ **(HiL)** and $\mathcal{D}_\mathbf{22}$ **(held-out)**. $\mathcal{D}_{21}$ is graded with the current calibrated model at temperature T and threshold τ to obtain abstentions and teacher corrections. We fine-tune with these corrections plus a replay buffer from $\mathcal{D}_0$ and then recalibrate. $\mathcal{D}_{22}$ is strictly reserved for final calibrated evaluation. Table 1 summarizes split sizes and distribution over M. In our data, $M = 4$ largely vanishes in recent exams, consistent with a shift towards longer, reasoning-heavy questions. This also motivates scale-aware replay sampling.

Label Rationale Since labels are ordinal and question scales differ (e.g., 0-4 vs. 0-10), we evaluate with Quadratic Weighted Kappa (QWK), which measures agreement while penalizing larger discrepancies more severely (off-by-2 is worse than off-by-1). QWK ranges from -1 (100% disagreement) to 1 (100% agreement). We report overall QWK on the full split and accepted set after abstentions. When scale-wise behavior is informative, we also report per-M QWK. This choice aligns the metric with grading semantics and real-world grading agreement in case of a conflict (Doewes et al., 2023).

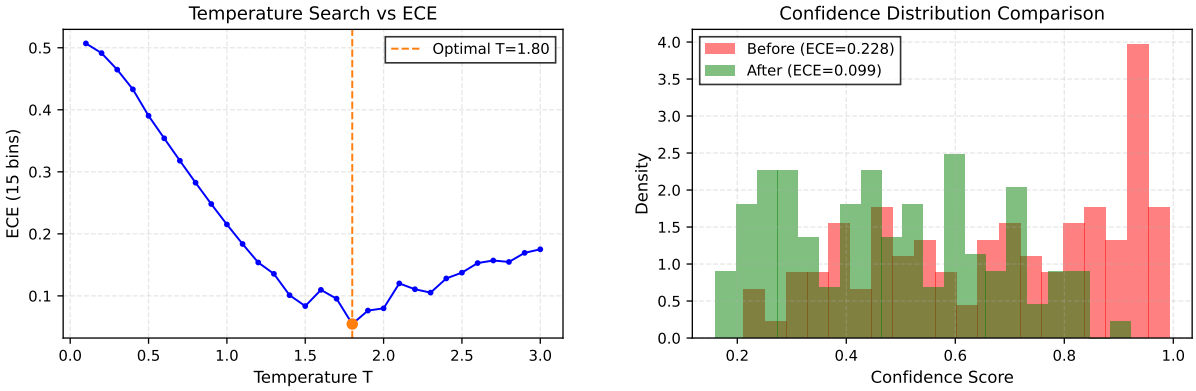

Figure 2: Calibration analysis and temperature scaling. (Left) Expected Calibration error (ECE) as a function of temperature T; the optimal T minimizes ECE. (Right) Confidence distribution before/after temperature scaling, showing reduced over-confidence and improved alignment with actual accuracy.

Stage	Split	T	τ	Coverage	Accepted	Abstained	QWK (acc.)
Uncalibrated	$\mathcal{D}_{21}$	1.0	–	0.700	132	57	0.600
Uncalibrated	$\mathcal{D}_{22}$	1.0	–	0.700	132	57	0.685
After Calibration	$\mathcal{D}_{21}$	1.8	0.340	0.624	118	71	**0.795**
After HiL round	$\mathcal{D}_{22}$	2.0	0.340	**0.772**	**146**	**43**	0.726

Table 2: Comparison of abstentions and accepted quality (measured as QWK) for uncalibrated vs. calibrated models with HiL.

4.2 Implementation

Architecture We use instruction-tuned chat models in the 3B-8B range to balance data scale and compute: Llama-3.2-3B-Instruct for the main results and Llama-3.1-8B-Instruct for ablation (Grattafiori et al., 2024). As discussed in Section 3.1, candidate scoring is not subject to any parameters, as we avoid an auxiliary classification head. Supervised training uses parameter-efficient fine-tuning implemented via PEFT/TRL with attention and MLP targets [q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj]. We implement data processing, training, and inference using the Hugging Face Transformers stack (Wolf et al., 2020). Unless stated otherwise, instruction-tuning defaults are: epochs=3, lr=2e-5, batch_size=2, grad_accum=16, cosine_scheduler, max_grad_norm=1.0, and max_length=2048. bF_16/fp16 plus gradient checkpointing are enabled only when needed to fit memory.

HiL Training HiL rounds fine-tune only LoRA adapters on the teacher-corrected abstentions and a replay buffer sampled by M (not predicted grade) to mirror the correction-scale distribution. This stabilizes adaptation across heterogeneous scoring scales. Updates are conservative by design to avoid overfitting: lr=5e-6, LoRA ($r = 16$, $\alpha = 32$, dropout $= 0.05$), and short schedules (1–2 warm epochs initially, then 1 epoch per HiL round) with small effective batches via gradient accumulation (4–16). After every HiL update, we recalibrate on $\mathcal{D}_1$ (temperature scaling to minimize 15-bin ECE) and select τ from the coverage-risk curve to meet a target coverage (e.g., 70%), then freeze τ for evaluation on $\mathcal{D}_{22}$. This setup works well in practice because it preserves the model's rubric-following generation, yields calibrated, comparable confidences for safe routing, and adapts to rubric drift using minimal additional parameters.

4.3 Evaluation

Grading Metrics We compute QWK on raw integer grades with the standard quadratic cost. We report overall accepted set QWK (auto-graded at τ), and per-scale QWK $(4/5/8/10)$ when scale-wise behavior matters. On the accepted set, we also summarize the off-by-1 and ≥ 2 misgrade rates to contextualize coverage.

Abstention Metrics Coverage is calculated as accepts/total and abstention as 1-coverage to quantify efficiency of HiL rounds. We additionally report acceptance composition (e.g., high-confidence vs

within-tolerance when applicable) to show how τ trades human effort for risk.

Unseen Exam Metrics For the final evaluation on $\mathcal{D}_{22}$ we follow: initial grading on $\mathcal{D}_{21}$ with calibrated T, threshold τ to obtain abstentions, post-HiL recalibration to T', τ' and evalaution on $\mathcal{D}_{22}$ with the updated model. We report first-half ($\mathcal{D}_{21}$) and second-half ($\mathcal{D}_{22}$) coverage/abstentions and accepted-QWK to quantify HiL impact.

5 Results

5.1 Calibration Outcomes

We first assess post-hoc calibration quality. A temperature sweep identifies the optimal scalar T on $\mathcal{D}_1$, minimizing 15-bin ECE while preserving precision. For Llama-3.2-3B-Instruct, $T \approx 1.8$ reduced ECE from 0.215 to 0.055 (-74.6%; see Figure 2). The confidence distributions shift from overconfident spikes to better-aligned probabilities after scaling. The reliability curves move closer to the identity line, corresponding to the ECE reductions (see Figure 3 in the Appendix B). We also observe the model is slightly under-confident, which is mitigated during the HiL and Post-HiL stages. These effects matter because $\hat{C} = \max_g p_g(T)$ is used for routing: better calibrated confidence means safer automation at a given coverage. Selecting $T = 1.8$ results in a lower ECE and a better confidence distribution, directly improving the quality and number of accepted decisions at any fixed coverage target.

5.2 Abstentions and Routing

With calibrated probabilities, we select an abstention threshold τ from coverage risk curves on $\mathcal{D}_1$ to achieve a target coverage. This makes the trade-off explicit: higher coverage reduces human effort, but may increase risk. The illustration of the coverage-risk curve in Figure 3 shows how we pick τ, and have a trade-off between quality of auto-grading and number of auto-accepts. The stage comparison in Table 2 aggregates both coverage/abstention counts and accepted-set QWK across splits $\mathcal{D}_{21}$ (After calibration) we accept 118/189, while on $\mathcal{D}_{22}$ (After HiL) we accept 146/189. The uncalibrated baselines are also presented for comparison.

We freeze τ within each evaluation phase but select it on $\mathcal{D}_1$ with a fixed coverage, and re-select it after HiL (post-recalibration). The uncalibrated model uses raw confidences (T=1.0) and calibrated uses temperature-scaled confidences (T=T'). Calibration does not change argmax labels, only confidences, and hence there will be differences in the samples that cross τ. On $\mathcal{D}_{21}$ (initial pass with pre-HiL calibrated model), **118** samples are accepted and **71** samples are abstained (71 corrections; simulated teacher). These, along with a replay buffer, drive the HiL update. At matched 70% coverage on $\mathcal{D}_{21}$, uncalibrated accepts/abstains at **132/57** (13 already-correct; 44 corrections required) versus calibrated **133/56** (12 already-correct; 44 corrections), indicating nearly identical behavior. On $\mathcal{D}_{22}$ before HiL, the calibrated model abstains 49 samples (12 already-correct; 37 corrections) versus uncalibrated model at 75 (18 already-correct; 57 corrections), yielding fewer, higher-value reviews with calibration. After HiL, on $\mathcal{D}_{22}$, routed samples drop to just 43 corrections out of 189 (with the re-selected τ), reflecting higher-yield routing post-adaptation, further highlighting the routed effort into gains on accepted decisions.

5.3 Grading Quality

Accepted-set QWK is high and stable across both stages. On $\mathcal{D}_{21}$ (calibrated model, T=1.8, τ=0.34), the accepted set attains QWK of **0.795** with **118** samples. After HiL and recalibration, on $\mathcal{D}_{22}$, the accepted-set reaches QWK of **0.726** with **146** samples. For context on $\mathcal{D}_{21}$ and $\mathcal{D}_{22}$ (uncalibrated model, T=1), the accepted-QWK is **0.600** and **0.685** respectively (see Table 2).

The per-M analysis in Table 3 shows that the accepted responses at M=10 dominate (D_{21}: n=59, QWK≈ 0.79; D_{22}: n=76, QWK$\approx$0.73) with comparatively low $\geq$ 2-point errors (0.169/0.355), indicating that longer answers admitted at τ=0.34 are typically of higher quality. On the contrary, the weakest performance comes from M=5, with higher severe errors, especially in D_{22} (QWK= 0.264, $\geq$ 2=0.530), pointing towards the medium difficulty level questions with neither too long or too short responses (usually, the consensus is that higher M requires clear, detailed, and explained answers) are quite risky to be auto-accepted. Such an analysis will help identify where human review is most valuable and should be prioritized to optimize overall grading reliability.

6 Discussion

Confidence calibration improves trustworthiness by aligning confidence with accuracy so that $\hat{C}$ is a

Stage	M	n	QWK	Off-by-1	≥ 2
$\mathcal{D}_{21}$	4	0	–	–	–
$\mathcal{D}_{21}$	5	54	0.447	0.204	0.463
$\mathcal{D}_{21}$	8	5	0.762	0.400	0.000
$\mathcal{D}_{21}$	10	59	0.788	0.153	0.169
$\mathcal{D}_{22}$	4	0	–	–	–
$\mathcal{D}_{22}$	5	66	0.264	0.136	0.530
$\mathcal{D}_{22}$	8	4	0.667	0.000	0.500
$\mathcal{D}_{22}$	10	76	0.725	0.066	0.355

Table 3: QWK Per-M and severe misgrade (≥ 2) rates on accepted sets at $\tau = 0.34$, quantifying agreement with human grades.

reliable routing signal. On $\mathcal{D}_1$, temperature scaling at $T \approx 1.8$ reduces ECE from 0.215 to 0.055 resulting in a smoother confidence distribution, which in turn supports clearer acceptance behavior. In deployment, comparing $\mathcal{D}_{21}$ vs $\mathcal{D}_{22}$ captures impact of the HiL round: coverage increases from 0.624 to 0.772, while there is a modest drop in QWK (-0.069). The baselines further show that calibration reduces the "already-correct" count in abstentions and directs focus on genuine errors.

Label efficiency and human correction efficiency are analyzed from the teacher correction statistics. Relative to the calibrated-only baseline ($\mathcal{D}_1$), a single HiL round increases auto-accepts at similar or better quality per unit supervision. Multi-seed reruns resulted in tiny fluctuations but stable calibrations, indicating robustness and dependability. Overall, the combination of teacher-forced scoring, temperature scaling, abstention policy, and HiL rounds delivers reliable and label-efficient automation.

Across $\mathcal{D}_{21}$ and $\mathcal{D}_{22}$, routed samples align with true errors, concentrating teacher efforts where they matter the most. Off-by-1 rates 0.186 on $\mathcal{D}_{21}$, and 0.096 on $\mathcal{D}_{22}$, compliment the accepted-QWK trends. However, the number of severe misgrades (≥ 2) rises (0.297 to 0.438), likely because the post-HiL confidence values admit more edge cases. Larger HiL rounds may mitigate this effect. Overall, the targeted corrections translate into a net QWK gain (0.05), demonstrating practicality in our system and motivating future work.

7 Conclusion

In this paper, we introduced a human-in-the-loop system for trustworthy, label-efficient automation of the short-answer grading task. Our system produces confidence estimates, ensuring that the model's self-assessed confidence aligns with actual grading reliability. By estimating confidence scores and setting a threshold to distinguish between confident and uncertain cases, the latter were rerouted for human evaluation, resulting in smoother confidence distributions and more reliable acceptance behavior. Moreover, a single human-in-the-loop iteration increased automatic grading coverage while preserving reliability and label efficiency, concentrating human effort on true errors. Across an unseen exam with scarce samples for final evaluation, the system ensured safe autonomy and transparent routing, highlighting the potential of deploying small, instruction-tuned models in parallel with teachers for courses with high cohort numbers. Future work will extend the system beyond the data mining domain to additional courses, exploring how grading performance generalizes across subjects with different rubrics, grading policies, and teacher styles. We also plan to extend our system by incorporating retrieval-augmented generation (RAG), utilizing lecture transcripts, textbooks, and course slides to provide context-aware feedback, thereby further enhancing transparency, adaptability, and pedagogical value. That, in addition to the HiL system, could then be deployed and distributed as an end-to-end, automatic, course-agnostic grading tool for teachers and institutions.

Limitations

Our corpus spans diverse topics related to the area of data mining, including classification and model evaluation, clustering, frequent itemset mining, and neural networks, hence providing a strong basis for generalization. At the same time, our experiments reflect a specific institutional context, and as part of future work, our intention is to explore additional datasets from other institutions. Nonetheless, the proposed methodology is course-agnostic and readily transferable to comparable educational settings. Moreover, the added inference cost with deterministic scoring is a trade-off for calibrated confidence values. Finally, although our HiL system mitigates demographic and inherent biases in pre-trained language models (Gallegos et al., 2024), we encourage end users and institutions to impose a clear acceptance threshold aligning to their desired risk tolerance.

References

Mustafa Abdul Salam, Mohamed Abd El-Fatah, and Naglaa Fathy Hassan. 2022. Automatic grading for arabic short answer questions using optimized deep learning model. *PLOS ONE*, 17(8):1–41.

Dishank Aggarwal, Pritam Sil, Bhaskaran Raman, and Pushpak Bhattacharyya. 2025. "i understand why i got this grade": Automatic short answer grading with feedback. *Preprint*, arXiv:2407.12818.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yucheng Chu, Peng He, Hang Li, Haoyu Han, Kaiqi Yang, Yu Xue, Tingting Li, Joseph Krajcik, and Jiliang Tang. 2025. Enhancing llm-based short answer grading with retrieval-augmented generation. *Preprint*, arXiv:2504.05276.

Emiliano del Gobbo, Alfonso Guarino, Barbara Cafarelli, and Luca Grilli. 2023. Gradeaid: a framework for automatic short answers grading in educational contexts—design, implementation and evaluation. In *Knowledge and Information Systems*.

Afrizal Doewes, Nughthoh Arfawi Kurdhi, and Akrati Saxena. 2023. Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 103–113, Bengaluru, India. International Educational Data Mining Society.

Ta Nguyen Binh Duong and Chai Yi Meng. 2024. Automatic grading of short answers using large language models in software engineering courses. In *2024 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–10.

Rafael Ferreira Mello, Cleon Pereira Junior, Luiz Rodrigues, Filipe Dwan Pereira, Luciano Cabral, Newarney Costa, Geber Ramalho, and Dragan Gasevic. 2025. Automatic short answer grading in the llm era: Does gpt-4 with prompt engineering beat traditional models? In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, page 93–103, New York, NY, USA. Association for Computing Machinery.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Alexey Gorbatovski and Sergey Kovalchuk. 2024. Reinforcement learning for question answering in programming domain using public community scoring as a human feedback. *Preprint*, arXiv:2401.10882.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Chris Impey, Matthew Wenger, Nikhil Garuda, Shahriar Golchin, and Sarah Stamer. 2025. Using large language models for automated grading of student writing about science. *International Journal of Artificial Intelligence in Education*.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. 2024. Understanding catastrophic forgetting in language models via implicit inference. In *The Twelfth International Conference on Learning Representations*.

Jimmy Ljungman, Vanessa Lislevand, John Pavlopoulos, Alexandra Farazouli, Zed Lee, Panagiotis Papapetrou, and Uno Fors. 2021. Automated grading of exam responses: An extensive classification benchmark. In *Discovery Science*, pages 3–18, Cham. Springer International Publishing.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786.

G. Mardini, D. Ivan, M. Quintero, G. Christian, N. Viloria, A. César, B. Percybrooks, S. Winston, N. Robles, and S. Heydy. 2024. A deep-learning-based grading system (asag) for reading comprehension assessment by using aphorisms as open-answer-questions. In *Education and Information Technologies*.

Seyedali Mirjalili, Seyed Mohammad Mirjalili, and Andrew Lewis. 2014. Grey wolf optimizer. *Advances in Engineering Software*, 69:46–61.

Nicolas Posocco and Antoine Bonnefoy. 2021. Estimating expected calibration errors. In *Artificial Neural Networks and Machine Learning – ICANN 2021*, pages 139–150, Cham. Springer International Publishing.

Arya Prabhudesai and Ta N. B. Duong. 2019. Automatic short answer grading using siamese bidirectional lstm based regression. In *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*, pages 1–6.

Rebecka Weegar and Peter Idestam-Almquist. 2024. Reducing workload in short answer grading using machine learning. In *International Journal of Artificial Intelligence in Education*.

Thomas Wolf, Lysandre Debut, Victor Sanh, ..., and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP 2020: System Demonstrations*, pages 38–45.

Juliette Woodrow, Chris Piech, and Sanmi Koyejo. 2025. Improving generative ai student feedback: Direct preference optimization with teachers in the loop. In *Proceedings of the 18th International Conference on Educational Data Mining*, pages 442–449, Palermo, Italy. International Educational Data Mining Society.

Johnathan Xie, Annie S Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. 2024. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18128–18138, Miami, Florida, USA. Association for Computational Linguistics.

Kamyar Zeinalipour, Mehak Mehak, Fatemeh Parsamotamed, Marco Maggini, and Marco Gori. 2025. Advancing student writing through automated syntax feedback. *Preprint*, arXiv:2501.07740.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2025. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

A Instruction Tuning

We performed instruction tuning experiments on both Llama-3.2-3B-Instruct and Llama-3.1-8B-Instruct models to understand the scalability and performance trade-offs. Training was conducted for 3 epochs using LoRA adapters ($r = 16$, $\alpha = 32$, dropout $= 0.05$) with a learning rate of 2e-4 and a batch size of 4. The 3B model required approximately 2 GPU hours, while the 8B model took around 4.5 GPU hours on NVIDIA RTX A5500 GPUs. Both models demonstrated robust convergence with very similar training and evaluation losses, though the 8B model achieved superior QWK performance (see Figure 4). Despite this performance advantage, we chose the 3B model for our main experiments due to the scale of our datasets ($\mathcal{D}_1$ and $\mathcal{D}_2$).

B Calibration Analysis

The reliability plot in Figure 3 shows that after calibration, the predicted confidence is much closer to the diagonal, indicating reduced overconfidence at higher grades. While the calibration is not perfect, we observe that the plot could be smoother, and achieve a higher confidence for higher grades (reflecting overall empirical accuracy), which we believe can be mitigated with additional data. The efficiency vs. quality tradeoff analysis in Figure 3 illustrates how varying τ, jointly changes the coverage and quality. It is estimated that for our target coverage of 70%, τ was determined to be 0.34.

C Reproducibility

The multi-seed reruns are stable (see Table 4): coverage varies by only $\pm 1.6\%$, and the accepted counts by ± 3 samples. The calibrated ECE stays within $\pm 0.05\%$, and we also note that the optimal T remains identical across seeds, indicating a deterministic calibration optimization.

Metric	Mean $\pm$ Std	Min	Max
Coverage	0.725 ± 0.016	0.709	0.746
ECE (before)	0.321 ± 0.007	0.313	0.330
ECE (after)	0.069 ± 0.005	0.062	0.074
Optimal T	2.0	2.0	2.0
Accepted	137 ± 3	134	141

Table 4: Reproducibility over three runs reporting mean$\pm$std, min, and max.

D Prompt and Examples

The structured prompt format (see Figure 5) ensures consistent parsing across heterogeneous question scales while maintaining the ordinal grading schema. The model learns to assign points by observing key concepts, and penalizing missing elements. We kept the prompt simple after observing that, for models of this size, lengthy prompts result in the model not obeying instructions (with max_new_tokens=96). Additionally, we observed that when the prompts contained too many instructions, the model forgets the specified *MaximumGrade* provided by the user.

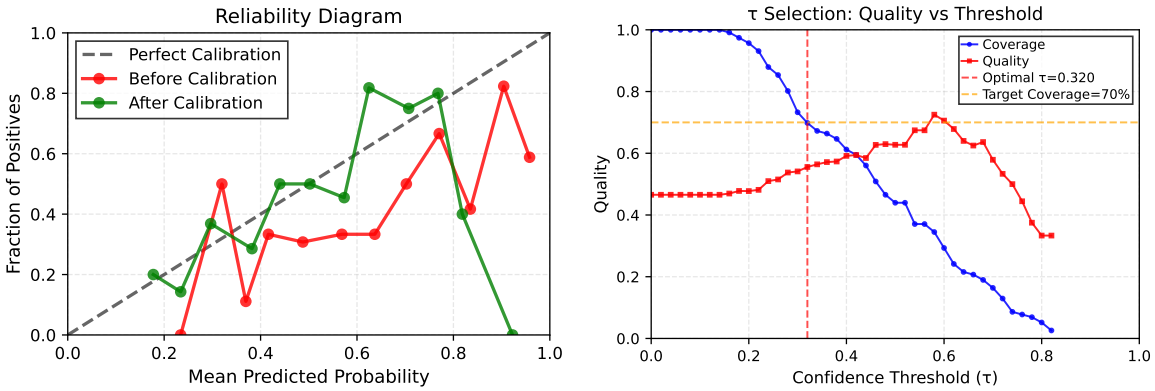

Figure 3: Reliability and coverage-risk plots. (Left) Calibrated vs. uncalibrated confidences, the diagonal indicating perfect calibration. (Right) Coverage-risk curve, with the threshold τ chosen to achieve 70% coverage. Quality is defined as 1-risk, i.e. the accuracy among auto-accepted samples at τ.

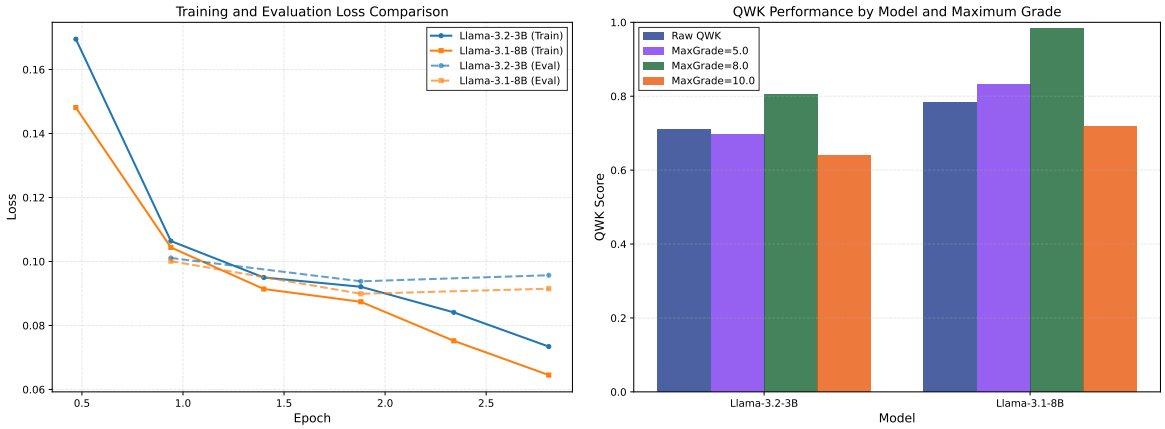

Figure 4: Training and validation performance comparison between Llama-3.2-3B-Instruct and Llama-3.1-8B-Instruct models. (Left) Training loss curves for both models reach a similar convergence, with the 8B model slightly lower. (Right) Validation QWK performance showing 8B model's significant advantage, indicating stronger agreement with human grades.

Prompt Template

System

You are an Automated Short Answer Grading (ASAG) system.
Return ONLY a strict JSON with keys: \"grade\" (int), \"max_grade\" (int)

User

Question: <Question>
Answer: <Response>
Target Scale: 0 to {<max_grade>}

Assistant

{{"grade": {g}, "max_grade": {mg}}}

Figure 5: Prompt template used for creating the instruction tuning dataset. The template specifies the student response and target score, and standardizes the model's grading instructions.