Automated Evaluation System of Descriptive Answers

Parthivi Singh[1] and Sakir Hussain[2]

1 - School of Computer Science and Engineering - Artificial Intelligence and Robotics
VIT Chennai, Kelambakkam - Vandalur Rd, Rajan Nagar, Chennai, Tamil Nadu 600127 - India

2 - School of Computer Science and Engineering
VIT Chennai, Kelambakkam - Vandalur Rd, Rajan Nagar, Chennai, Tamil Nadu 600127 - India

Abstract. This project introduces an automated assessment system for descriptive responses from students, combining Large Language Models (LLMs) with knowledge-based retrieval and anticipatory reasoning. The system integrates Graph-Based Retrieval-Augmented Generation (Graph RAG) with a Proactive Chain-of-Thought (ProCoT) evaluation framework to evaluate responses with minimal human involvement. A knowledge graph tailored for theoretical computer science content facilitates context-sensitive question creation and assessment. The dataset includes ten assignments and two exams from an introductory computer science course at the University of North Texas, with student responses assessed by two annotators on a correctness scale ranging from 0 to 5. The assessment procedure involves a Non-Collaborative Pre-processing, Clarification and Target-Guided dialogue , facilitating both deduction and the creation of marks. The system employs DeepSeek-Coder 32B and Mixtral 8x7B models through Ollama for reasoning and generating answers, utilizes BGE-M3 embeddings alongside FAISS for semantic search, and features a Post-Hoc Softener Model that refines scores based on SBERT similarity, TF-IDF scores, answer length, and keyword overlap. Test outcomes based on 300 samples reveal an 89.72% correlation with human ratings, yielding a Mean Absolute Error (MAE) of 0.53 and a Root Mean Squared Error (RMSE) of 0.79, validating the system's dependability and performance in accordance with human evaluation.

1 Introduction

In modern educational ecosystems, student assessment is a critical component that directly influences academic progression, career prospects, and overall learner development. Among the various modes of assessment, the evaluation of descriptive or long-form answers plays a particularly vital role. Unlike objective-type evaluations, descriptive assessments enable educators to measure a student's conceptual understanding, analytical skills, and ability to articulate ideas coherently. They serve as an effective tool for gauging higher-order thinking skills such as critical analysis, reasoning, and the application of domain-specific knowledge.

Despite their pedagogical importance, the evaluation of descriptive answers remains a labor- intensive and subjective task. Manual assessment by human evaluators is often plagued by inconsistencies arising from evaluator fatigue, personal biases, and variability in interpretation. Moreover, with the rise of large-scale educational programs, digital learning platforms, and hybrid classroom environments, the scalability of manual evaluation has become a pressing concern. Institutions are increasingly seeking automated, accurate, unbiased assessment solutions and are capable of providing timely and constructive feedback to students. While the integration of Artificial Intelligence (AI) in educational technology has seen significant advancements in recent years, the automation of subjective answer evaluation remains a challenging problem. Unlike objective assessments, descriptive answers exhibit considerable linguistic variability, diverse expression styles, and contextual dependencies, making their evaluation complex and non-trivial. Existing automated grading systems often fall short in handling these nuances, leading to fairness issues, lack of transparency, and student dissatisfaction.

This project addresses these challenges by proposing a robust, scalable, and transparent system for the automated evaluation of descriptive answers. The proposed system leverages advancements in Natural Language Processing (NLP), specifically integrating techniques such as knowledge-grounded language generation, structured reasoning, and post-hoc score adjustment to mimic human evaluation behavior while ensuring consistency and fairness. The core objectives of this work are to minimize subjectivity and bias in descriptive answer evaluation, provide explainable and transparent grading outcomes, ensure scalability and efficiency in large-scale academic settings, and deliver scores that closely align with human evaluator judgments.

The proposed system is designed to not only automate the evaluation process but also provide constructive feedback to students, thereby enhancing the learning experience and fostering continuous academic improvement. This work contributes to the growing body of research in AI-assisted education by addressing one of its most challenging and underexplored areas — the fair and explainable evaluation of long-form, descriptive student responses

2 Related Works

2.1 Evaluating Descriptive Answers Using Large Language Models

The automated assessment of descriptive responses has gained increasing attention with the advancement of Large Language Models (LLMs), particularly in contexts requiring higher-order reasoning and conceptual understanding. Traditional grading approaches—keyword matching, TF-IDF similarity, and rule-based systems—have been criticized for failing to accurately evaluate open-ended responses, especially in fields demanding critical thinking [5].

To overcome these limitations, multiple frameworks have been proposed. One of the foundational studies in this space introduces reference-based grading as one of three LLM judge strategies, focusing on aligning model scoring with human preferences using rubrics and reference responses [14]. This aligns with the evaluation objectives of our system, which emphasizes rubric-based and reference-grounded grading for descriptive answers.

2.2 Proactive Reasoning and Dialogue-Based Evaluation

The Proactive Chain-of-Thought (ProCoT) framework was proposed to structure LLM-based evaluation through multi-turn interactions—namely Clarification Dialogue, Target-Guided Dialogue, and Non-Collaborative Dialogue. This model improves LLM performance by enabling planning, refinement, and reasoning steps rather than relying on single-shot outputs [3]. Our system incorporates ProCoT to conduct deductive and constructive grading phases, enabling both penalization and award of marks based on thematic gaps and conceptual completeness.

In a similar spirit, the RCOT (Reverse Chain-of-Thought) method deconstructs generated answers to verify factual consistency and reasoning correctness by comparing derived and original question conditions [13]. While RCOT focuses on post-hoc error detection, ProCoT emphasizes proactive scoring, making them complementary approaches for future hybrid evaluation frameworks.

2.3 Grading Frameworks and LLM Ensembles

Several grading systems explore LLM-based feedback with structured evaluation mechanisms. CodEv utilizes LLM ensembles and Chain-of-Thought reasoning for code assessment, yielding consistent scoring and constructive feedback. It emphasizes scalability through distillation to smaller models, a strategy directly relevant to future directions in descriptive answer evaluation [12]. Similarly, GradeOpt introduces multi-agent collaboration between LLMs (Grader, Reflector, Refiner) for refining rubrics and improving grading consistency, especially under out-of-distribution scenarios [2]. These insights reinforce the benefits of reflective multi-agent grading pipelines.

Meanwhile, SteLLA and IntelliGrader present structured grading pipelines by transforming rubrics into question-answer form [9] and combining linguistic features with regression models for short answer grading [11], respectively. Both approaches advocate for human-aligned scoring and feedback synthesis, core principles in our project's architecture.

2.4 Benchmarking Models and Model Selection

A recent benchmarking study evaluates models like GPT-3.5, GPT-4, Claude-3, and Mistral-Large using RAG for short answer grading. It highlights model-specific tendencies: Claude-3 being more cautious, GPT-3.5 being radical, and Mistral-Large being most consistent [6]. These findings support our design choice to use Mixtral 8x7B for answer generation and scoring.

Additionally, the EngSAF and MMSAF datasets from IIT Bombay introduced feedback-enhanced ASAG tasks across both textual and multimodal formats. These studies reported that models like Mistral-7B performed best in generating accurate, structured feedback, especially for concept-heavy responses ([1], [10]). The incorporation of feedback in these systems further motivates our Post-Hoc Softener Model, which adjusts scores based on content relevance and structure.

2.5 Future Directions and Research Gaps

Recurring challenges across studies include bias mitigation, rubric subjectivity, and hallucination control. Suggested solutions include standardized rubrics, hybrid models combining general and domain-specific LLMs, and ensemble methods for feedback verification [12]. In addition, step-by-step distillation of large models into smaller, task-specific variants was proposed to reduce computational overhead while retaining performance [12].

Our system responds to these gaps by using domain-specific knowledge graphs, structured dialogue-based grading, and score refinement modules—all aligned with human rubrics and feedback standards.

3 Proposed Methodology

The suggested approach describes a modular, multi-step process aimed at automating the assessment of descriptive student responses with great reliability and consistency with human grading standards. The process includes five main phases: context-sensitive optimal response generation, student answer preparation, rubric-based assessment through Proactive Chain-of-Thought (ProCoT), organized output verification, and score modification after evaluation.

3.1 Ideal Answer Generation via Graph-Based Retrieval-Augmented Generation (Graph RAG)

The assessment procedure starts by creating a reference (ideal) response through a domain-specific retrieval system. The academic source corpus—consisting of textbook content and selected theoretical materials—is processed, refined, and divided into logical segments. Every segment is embedded with the BGE-M3 embedding model through Ollama's API and indexed with FAISS for quick similarity retrieval.

A knowledge graph is subsequently created using NetworkX, with each chunk represented as a node and edges formed based on semantic similarity. When a question is posed, the system acquires the top-k pertinent chunks using FAISS and then broadens the context by exploring adjacent nodes in the graph. The combined context is subsequently sent to a Mixtral 8x7B model to produce a detailed, syllabus-based optimal response.

3.2 Student Answer Preprocessing via Non-Collaborative Filtering

Prior to evaluation, student responses are screened to eliminate irrelevant or nonsensical parts through a Rolling Context Filtering process. This module calculates semantic similarity with SBERT, measures lexical overlap through TF-IDF, and performs relevance classification using a zero-shot classifier based on BART. Sentences are kept only if a minimum of two out of the three models consider them relevant based on a tolerance-guided voting system. This guarantees that the following scoring stage assesses solely relevant content, enhancing scoring accuracy.

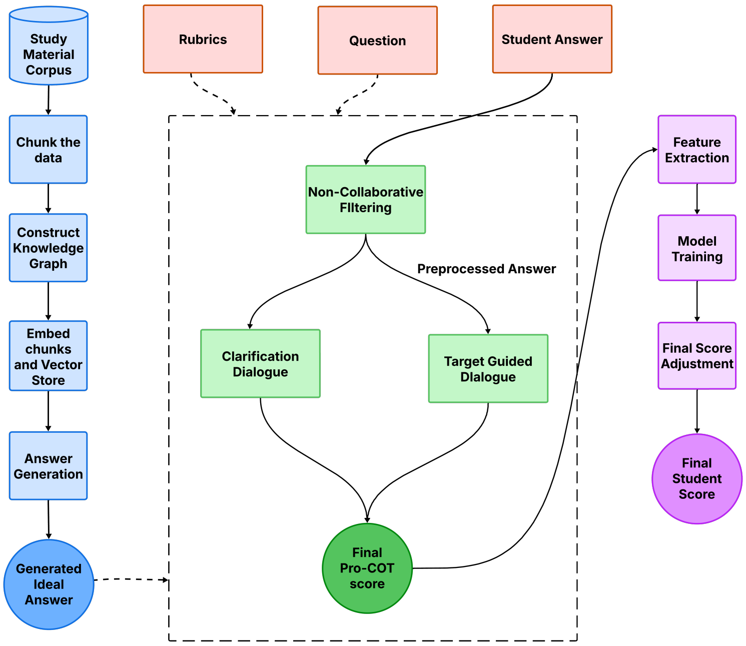

Fig. 1: Architecture Diagram for Proposed System

3.3 Rubric-Driven Evaluation using ProCoT Reasoning

Filtered responses are assessed according to rubric items utilizing the Proactive Chain-of-Thought (ProCoT) framework. For every rubric, two strategies based on dialogue reasoning are utilized:

1. Clarification Dialogue: Identifies gaps or unclear parts in the student's answers, mimicking the way a human evaluator recognizes absent rubric components.
2. Target-Directed Dialogue: Assesses the number of thematic changes needed to turn the student's reply into the optimal response, using both SBERT and TF-IDF similarity measures to evaluate alignment.

Every dialogue stage lasts for three iterative turns and produces a structured JSON output that includes the evaluator's reasoning, the action performed, and the final adjusted score (normalized from 0 to 1). These results are upheld and verified through Guardrails, which enforce rigid schema adherence through RAIL templates to guarantee output clarity and traceability.

3.4 Structured Output Enforcement via Guardrails

To maintain uniformity and structural integrity of all produced evaluations, the system incorporates the Guardrails AI framework. This schema mandates the inclusion of essential evaluation elements like the selected evaluation method, reasoning, actions, feedback, and score. Any improperly formatted outputs are automatically discarded or fixed before further processing. This system guarantees that every response is readable by machines, dependable, and can be tracked.

3.5 Post-Evaluation Score Adjustment via Softener Model

Even with a structured scoring system, early results indicated a consistent bias towards under-evaluating student responses in comparison to human evaluators. To tackle this, a Post-Hoc Softener Model was created utilizing a Gradient Boosting Regressor trained on samples evaluated by humans. The model receives four features as input: raw ProCoT score, SBERT similarity, TF-IDF similarity, and length of the answer. It forecasts a milder score that more accurately represents human tolerance. The ultimate score is calculated as a weighted combination of ProCoT and adjusted scores, fine-tuned to reduce MAE and RMSE while enhancing correlation with human evaluations.

4 Performance and Observations

4.1 Dataset and Preparation

The assessment pipeline was evaluated using the UNT Computer Science Short Answer Dataset (Version 2.0), a frequently referenced dataset for research on grading open-ended academic responses. The collection consists of:

- 10 tasks (4–7 questions per task)
- 2 tests (10 questions per test)
- More than 2,200 student responses in total
- For every question: the prompt, a model response crafted by an instructor, and two scores given by humans on a 0–5 integer scale.

Nonetheless, the dataset lacks any rubrics, which are crucial for evaluation based on criteria. To tackle this issue, a rubric creation module was designed utilizing a language model influenced by Guardrails schemas. For every [question, ideal answer] combination, the system produced 3–5 rubric elements, each embodying a separate idea, guaranteeing the overall total aligned with the initial 5-mark framework. These rubrics were later employed to assess student responses in a structured and understandable manner.

4.2 Evaluation Setup

To assess the effectiveness of the suggested auto-evaluation pipeline, a sample of 1,000 student responses was chosen from the dataset. Every response was handled utilizing the entire evaluation pipeline, which comprises:

1. Non-Collaborative Filtering for eliminating unimportant content
2. Proactive Chain-of-Thought (ProCoT) assessment based on rubric criteria (Clarification + Target-Oriented Dialogues)
3. Validated output structure utilizing Guardrails
4. Post-Hoc Softener Model for aligning scores to mimic human behavior

The experiments were conducted on one NVIDIA H100 GPU. The main assessment model employed for dialogue reasoning was DeepSeek-Coder R1 32B, run locally through Ollama. Semantic similarity embeddings were produced using bge-m3 through FAISS for retrieval and SBERT for scoring similarity.

4.3 Evaluation Metrics

To quantitatively evaluate the effectiveness of the automated assessment pipeline, various commonly recognized scoring metrics were employed to contrast the scores produced by the system with the human-annotated reference. These metrics offer a thorough insight into both the precision and conformity of the system's results with human expectations:

- **Mean Absolute Error (MAE)**: This metric calculates the average absolute difference between the system's predicted score and the human-assigned score. A lower MAE indicates a closer approximation to human evaluation.
$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{1}$$
- **Root Mean Squared Error (RMSE)**: RMSE penalizes larger errors more heavily than MAE by squaring the differences before averaging and then taking the square root. It reflects the magnitude of deviation from human scores.
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{2}$$
- **Pearson Correlation Coefficient (PCC)**: This metric measures the linear correlation between the predicted scores and the human scores. Values close to 1 indicate strong positive correlation.
$$\text{PCC} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \cdot \sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{3}$$
- **Score Alignment Distribution**: Apart from continuous metrics, we also track categorical score behaviors including:

  1. Perfect Matches: Exact matches between system and human scores.

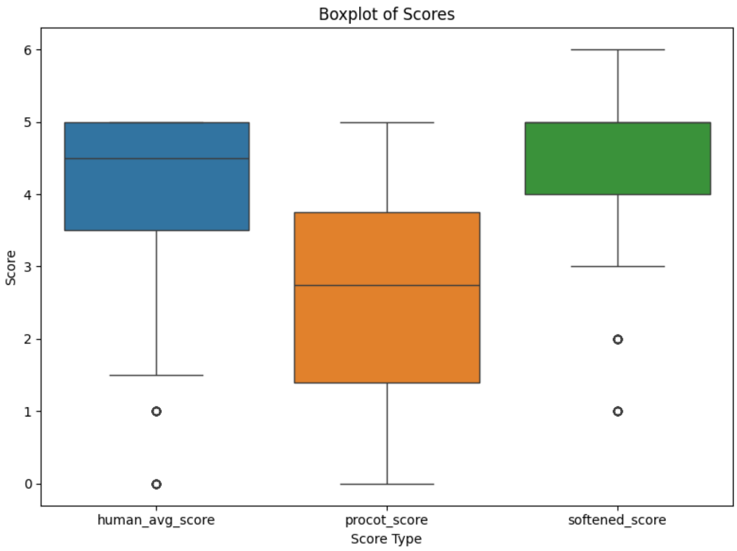

Fig. 2: Boxplot to compare scores given directly by the ProCOT module and the Softner.

2. Underscored Cases: Instances where the system scored lower than human annotators.
3. Overscored Cases: Instances where the system scored higher than human annotators.

These metrics allow for a rigorous, multi-dimensional analysis of the system's grading behavior, highlighting not just raw accuracy but also tendencies toward conservative or generous scoring.

4.4 Observations: Systematic Underscoring and Correction

Early findings from the ProCoT assessment showed a consistent pattern of underestimation:

- Just 14.50% of ProCoT scores aligned perfectly with the human average.
- 83.30% of student responses were rated lower compared to human reviewers, while a 0.71 Person Coefficient showed a good alignment with human grading patterns
- The Mean Absolute Error (MAE) was 1.53, while RMSE attained 1.86—showing substantial variation in raw scores.

This phenomenon was linked to the ProCoT system's rigid, deductive scoring approach, which does not exhibit the flexibility often seen in human evaluation—especially for answers that are partially correct or stylistically different.

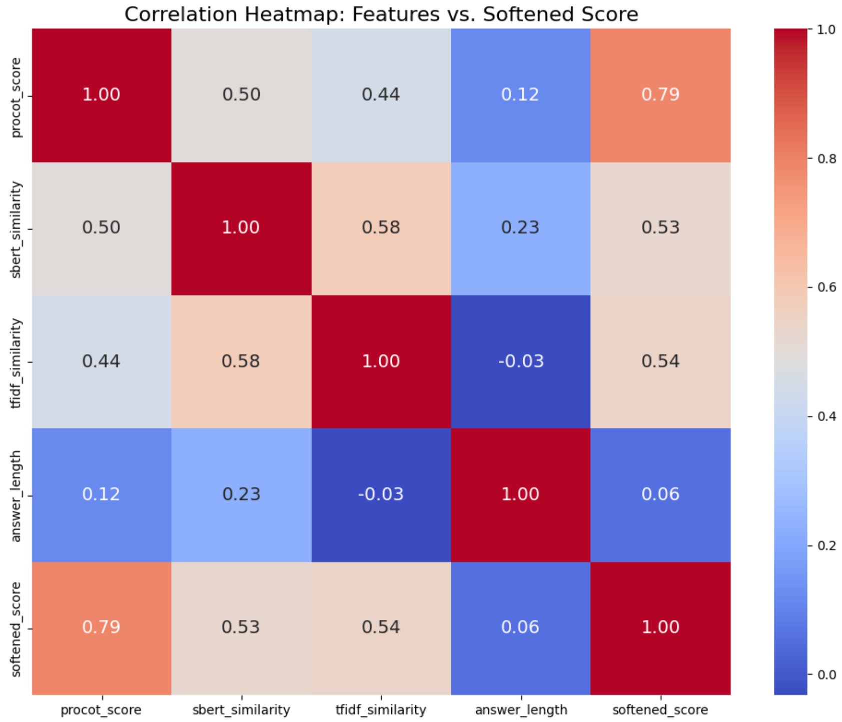

Fig. 3: Heatmap to show co-relation between chosen features for the Gradient Booster

To tackle this issue, a Post-Hoc Softener Model was implemented. This model, developed with Gradient Boosting Regression, forecasted human-aligned scores by utilizing these features:

1. Unprocessed ProCoT score

2. SBERT resemblance between student response and perfect answer

3. TF-IDF similarity

4. Response length

The adjusted scores were created by combining the softened scores with the ProCoT outputs. This adjustment technique significantly enhanced score alignment:

These findings confirm the necessity for a post-evaluation modification stage in AI-driven grading systems. The softener model not only mitigated the severity of raw ProCoT scores but also aligned the system's scoring behavior much more closely with that of human assessors.

Metric	ProCoT Score	Final Adjusted Score
MAE	1.5312	0.5325
RMSE	1.8658	0.7942
PCC	0.7091	0.8261
Underscored Responses	83.30%	7.30%
Perfect Matches with Human Score	14.50%	44.10%

Table 1: Comparison of evaluation metrics between ProCoT scoring and the Final Adjusted (Softened) scores.

5 Summary

In conclusion, this study introduces a modular, knowledge-based system for the automated assessment of descriptive student responses, tackling the fundamental issues of subjectivity, scalability, and equity in academic evaluation. The system incorporates a Graph-Based Retrieval-Augmented Generation (Graph RAG) model for creating ideal answers aligned with the syllabus, a Non-Collaborative Filtering component to filter out unrelated student inputs, and a rubric-guided Proactive Chain-of-Thought (ProCoT) assessment engine that emulates human-like multi-step reasoning. All outputs undergo thorough validation with Guardrails to guarantee organized, understandable evaluation documentation. To tackle the systemic underscoring noted during early evaluations, a Post-Hoc Softener Model was integrated, utilizing elements like semantic similarity, lexical alignment, and answer length to enhance the final score. The evaluation of the system was conducted using 1,000 responses from the UNT Short Answer Dataset, analyzed on a DeepSeek-Coder R1 32B model running on H100 GPU. The findings revealed a significant decrease in error—lowering MAE from 1.53 to 0.53—and a notable enhancement in Pearson correlation from 0.71 to 0.83, showing a strong correlation with human assessment. These results confirm the efficacy of the suggested pipeline in generating precise, understandable, and human-centered evaluations on a large scale, while establishing the groundwork for future progress in automated educational assessment.

6 Conclusion and Future Work

Although the automated evaluation system created in this project tackles significant issues related to descriptive answer assessment, there are still many promising avenues for future improvement and wider applicability. These instructions highlight both the constraints of the existing implementation and the system's capability to cater to a wider range of educational environments.

A key domain for growth includes broadening the system beyond its present emphasis on theoretical computer science. While the current implementation has been verified in this area, the modular structure allows for application across various academic fields. Subsequent versions might integrate field-specific resources and grading frameworks from areas like social sciences, humanities, and engineering—thus facilitating implementation in a broader array of institutional environments.

Another important opportunity exists in promoting multimodal assessment. Currently, the system assesses solely textual replies. Nevertheless, in various academic situations, students enhance their responses with diagrams, charts, flowcharts, or symbolic images that express essential meaning. Combining computer vision methods with multimodal reasoning frameworks would enable the system to understand and assess visual components in conjunction with text particularly beneficial in STEM and design-focused fields.

Enhancements to the Non-Collaborative Filtering module also offer a path for advancement. The existing design uses set thresholds and pre-trained relevance classifiers to eliminate irrelevant or unrelated content. Upcoming improvements might integrate adaptive filtering techniques that flexibly modify sensitivity according to question difficulty, topic area, or detected writing habits of students, possibly leveraging insights from earlier assessments.

Enhancements to the Post-Hoc Softener Model also deserve investigation. Although the existing model relies on a predetermined collection of linguistic features and gradient boosting regression, upcoming iterations might include more sophisticated metrics like syntactic complexity, coherence, and discourse structure. Furthermore, broadening the training dataset to incorporate a wider range of human-evaluated responses would enhance generalizability and minimize possible bias.

From a deployment perspective, connecting the system with Learning Management Systems (LMS) is a sensible next move. By offering standardized APIs and user-friendly interfaces, the system can be smoothly integrated into academic processes, enabling instructors to automate large-scale grading while maintaining oversight and transparency.

A notably significant improvement includes the implementation of personalized evaluator fine-tuning via Parameter-Efficient Fine-Tuning (PEFT) approaches like LoRA or adapter-based strategies. This would enable individual instructors to adjust evaluator profiles based on a limited set of annotated examples, allowing the system to align with personal grading approaches while preserving rubric integrity and fairness. These lightweight adapters can be applied dynamically to the base model, enabling instructor-specific grading while maintaining scalability.

Ultimately, the system would gain from the ongoing incorporation of improvements in language model designs, organized prompting, and output management systems. Integrating advanced LLMs, refined Guardrails schemas, and dynamic prompt optimization techniques will further boost the system's precision, clarity, and resilience.

In summary, the methods and framework developed in this project offer a strong basis for scalable, rubric-based evaluation of descriptive responses. The future paths outlined here provide a distinct guide for enhancing the system's adaptability, equity, and effectiveness—significantly aiding the advancement of smart educational evaluation instruments.

References

[1] Dishank Aggarwal, Pushpak Bhattacharyya, and Bhaskaran Raman. " i understand why i got this grade": Automatic short answer grading with feedback. *arXiv preprint arXiv:2407.12818*, 2024.

[2] Yucheng Chu, Hang Li, Kaiqi Yang, Harry Shomer, Hui Liu, Yasemin Copur-Gencturk, and Jiliang Tang. A llm-powered automatic grading framework with human-level guidelines optimization. *arXiv preprint arXiv:2410.02165*, 2024.

[3] Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. *arXiv preprint arXiv:2305.13626*, 2023.

[4] Dong Dong and Yue Liang. Grading programming assignments by summarization. In *Proceedings of the ACM Turing Award Celebration Conference-China 2024*, pages 53–58, 2024.

[5] Rujun Gao, Hillary E Merzdorf, Saira Anwar, M Cynthia Hipwell, and Arun R Srinivasa. Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*, 6:100206, 2024.

[6] Jussi S Jauhiainen and AgustÃn Garagorry Guerra. Evaluating students' open-ended written responses with llms: Using the rag framework for gpt-3.5, gpt-4, claude-3, and mistral-large. *arXiv preprint arXiv:2405.05444*, 2024.

[7] Jordan K Matelsky, Felipe Parodi, Tony Liu, Richard D Lange, and Konrad P Kording. A large language model-assisted education tool to provide feedback on open-ended responses. *arXiv preprint arXiv:2308.02439*, 2023.

[8] Ifeanyi G Ndukwe, Chukwudi E Amadi, Larian M Nkomo, and Ben K Daniel. Automatic grading system using sentence-bert network. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*, pages 224–227. Springer, 2020.

[9] Hefei Qiu, Brian White, Ashley Ding, Reinaldo Costa, Ali Hachem, Wei Ding, and Ping Chen. Stella: A structured grading system using llms with rag. In *2024 IEEE International Conference on Big Data (BigData)*, pages 8154–8163. IEEE, 2024.

[10] Pritam Sil, Bhaskaran Raman, and Pushpak Bhattacharyya. " did my figure do justice to the answer?": Towards multimodal short answer grading with feedback (mmsaf). *arXiv preprint arXiv:2412.19755*, 2024.

[11] Paradesi Sree Lakshmi, Jay B Simha, and Rajeev Ranjan. Intelligrader: A framework for automatic short answer grading, inconsistency check and feedback in educational context-conception, implementation and evaluation. *Karbala International Journal of Modern Science*, 10(3):9, 2024.

[12] En-Qi Tseng, Pei-Cing Huang, Chan Hsu, Peng-Yi Wu, Chan-Tung Ku, and Yihuang Kang. Codev: An automated grading framework leveraging large language models for consistent and constructive feedback. In *2024 IEEE International Conference on Big Data (BigData)*, pages 5442–5449. IEEE, 2024.

[13] Tianci Xue, Ziqi Wang, Zhenhailong Wang, Chi Han, Pengfei Yu, and Heng Ji. Rcot: Detecting and rectifying factual inconsistency in reasoning by reversing chain-of-thought. *arXiv preprint arXiv:2305.11499*, 2023.

[14] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.