

# Bi-Contextual Retrieval Augmented Generation (RAG) for Automatic Descriptive Answer Grading

**Asma Amjad**

*Department of Information Technology,  
The Islamia University of Bahawalpur Bahawalpur,  
Pakistan*

Asma.amjad@iub.edu.pk

**Malik Muhammad Saad Missen**

*Department Software Engineering,  
The Islamia University of Bahawalpur, Bahawalpur,  
Pakistan*

saad.missen@gmail.com

**Muzamil Malik**

*Department of Computer Science,  
Hamdard University, Islamabad,  
Pakistan.*

saad.missen@gmail.com

**Hassan Taimour Khan**

*Department of Technology Management,  
The Islamia University of Bahawalpur  
Pakistan.*

Hassan.taimour@iub.edu.pk

**Corresponding Author:** Malik Muhammad Saad Missen

**Copyright** © 2026 Asma Amjad, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Abstract

Automatic Short Answer Grading (ASAG) is a well-known research task in the field of natural language processing (NLP). Its major purpose is to automatically grade descriptive answers of the students by keeping automatic grading consistent with the evaluation of human graders. Recent developments in Large Language Models (LLMs) have demonstrated a greatly enhanced performance in automated grading; however, the generalizability of the models and accuracy is still quite low because of the absence of dataset-specific grounding. We present EDURAG, a Retrieval-Augmented Generation (RAG) based model to improve contextualization of the LLM-based grading with exemplar-based grading and extra knowledge as generated by QFKE (Question Focused Knowledge Extraction) module. The proposed QFKE module provides extra layer of contextuality for the EDURAG. In contrast to conventional supervised methods, EDURAG does not need model fine-tuning. The suggested framework is tested against the ASAG2024 benchmark that consolidates seven short-answer grading datasets across various domains, educational levels, and grading scales. The benchmark protocol of measuring performance is weighted Root Mean Square Error (wRMSE). The experimental findings show that dual contextuality provided by EDURAG enhances the accuracy of grading significantly when compared to vanilla LLM grading. The ablation study also confirms the significance of dual contextuality provided by EDURAG. AI-

though promising results (almost 15% improvement) have been achieved during experiments, there is still a discrepancy between human grading performance and merit, indicating the potential of hybrid human-AI grading systems. The results indicate that retrieval-enhanced LLMs offer a scalable and generalizable direction for automated assessment.

**Keywords:** Automatic grading, LLM, Short answers, Machine learning.

## 1. INTRODUCTION

Evaluation of students learning is one of the most crucial phases in education. Traditional evaluation involves manual grading of paper-based answer sheets. While many have consulted automatic grading systems, but their use remains limited to grading of multiple-choice questions or fill-in-the-blanks where the answers are short and easy to compare with a fixed key [1]. These systems are not sufficient for the evaluation of descriptive answers, where students write explanations in natural language. Descriptive answers are common in science, history, and literature, and they give deeper insight into student understanding. But they are also difficult to grade because the same idea can be expressed in many ways, with different words, order, or level of detail [2].

Manual grading by teachers is still considered the most reliable method for descriptive answers. However, it requires a large amount of time and effort, especially in large classes or online education environments. Grading inconsistency is one of the major issues students face in manual grading [3]. All these challenges motivate research community to look for better alternatives.

Emergence of Large Language Models (LLMs) [4] can be considered an excellent opportunity for the development of effective automatic grading systems. It has been experienced that LLMs understand natural language to the extent where it becomes challenging to differentiate between computers and human beings. Several studies have reported that LLMs can reach moderate agreement with human graders on short-answer datasets [5, 6]. However, LLMs still face some issues when employed for this task. For example, LLMs often lack access to the specific knowledge [7] that is needed to evaluate domain-heavy subjects like science or history. However, this issue can be solved if domain-specific information can be provided to LLMs. This is where the approach of Retrieval Augmented Generation (RAG) comes to aid us.

### 1.1 Retrieval Augmented Generation (RAG)

To overcome the default issue of LLMs, a promising approach is Retrieval-Augmented Generation (RAG). RAG combines two parts: a retrieval system and a language model. The retrieval system searches for relevant knowledge or examples from external sources, and the language model then uses this information to produce a more accurate and context-aware output [8] as shown in FIGURE 1. In the case of automatic grading, retrieval can provide curriculum material, rubrics, or exemplar student responses, and the LLM can then judge the new student answer against this evidence. This approach helps reduce hallucination, improves fairness, and aligns the grading more closely with teacher expectations [9, 10].

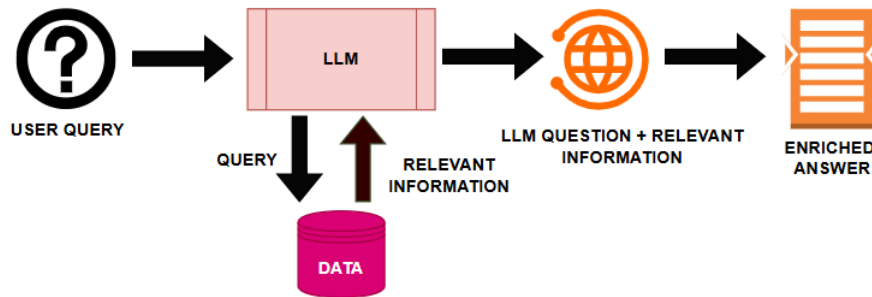


Figure 1: Overview of the RAG process

In this paper, we present EDURAG, a framework that applies RAG to the task of descriptive answer grading. The proposed model comes with unique approach of providing dual-context to RAG to grade descriptive answers. In addition to this, it also highlights the supporting evidence and generates short formative feedback. The model is evaluated using 07 benchmark datasets included in ASAG2024 and a fair compare with potential baselines is provided. The results demonstrate the effectiveness of the proposed RAG-based model.

## 1.2 Contributions

This work makes the following key contributions:

- We propose EDURAG, a retrieval-augmented grading framework that enhances Large Language Model (LLM)-based short-answer assessment by incorporating dual context to RAG. One context is extracted from the data itself while other one is generated using QFKE module which retrieves question focused knowledge to be fed to the RAG.
- We introduce a unique approach of providing dual context to our model for ASAG task. The QFKE module utilizes the potential of LLMs to generate question focused extra knowledge for the RAG which uses it to grade the answers.
- We provide a comprehensive evaluation of retrieval-augmented LLM grading on the ASAG2024 benchmark, which combines seven short-answer grading datasets into a unified evaluation framework.
- We conduct ablation analysis to quantify the individual and combined impact of different components of the proposed system.

Unlike existing RAG-based grading approaches that rely solely on retrieved exemplars or reference materials, this work introduces a dual-context augmentation strategy, where both retrieved grading patterns and generated conceptual knowledge are jointly leveraged. This enables the model to better handle semantic variability in student responses while aligning predictions with human grading behavior.

### 1.3 Structure of the Paper

This is how this paper is structured: In section 2, we provide a highlight of the related work. The architecture of the proposed model is well described in section 3 and in section 4 we introduce our proposed model EDURAG. The experimental setup is described in section 5 while detailed results are presented in section 6. Any risk factors and implications are discussed in section 7 along with deployment details. The limitations of the work are given in section 8 and we conclude the paper in section 9 along with highlights of the future work.

## 2. RELATED WORK

A significant amount of work has been published on automatic grading of student answers during the previous decades. It involves approaches using simple rule-based techniques to modern neural networks and large language models (LLMs). In the following sub-sections, we discuss some of these works grouped by their methodology.

### 2.1 Rule-Based and Semantic Similarity Approaches

Early work in automatic grading mainly relied on handcrafted rules and semantic similarity measures. The Intelligent Essay Assessor (IEA) applied Latent Semantic Analysis (LSA) to capture semantic overlap between student essays and reference texts, achieving performance comparable to human graders [11]. Similarly, c-rater [12] used pattern matching and hand-coded linguistic rules to grade short answers in science and reading comprehension tasks. These techniques were good in small-scale areas but lacked flexibility and could not respond effectively to open-ended or creative responses.

### 2.2 Feature Engineering with Machine Learning

The next wave of systems focused on engineered features (e.g., lexical overlap, syntactic similarity, n-grams) combined with traditional machine learning models. Mohler, Mihalcea, and Mihalcea (2011) [13] have shown that the use of semantic similarity features plays a major role in enhancing automatic short answer grading. Osaka et al. (2025) [14], proposed a workflow for automated short-answer scoring based on active and deep learning for large datasets. These approaches improved over purely rule-based systems but required labor-intensive feature design and still struggled with unseen topics. TABLE 1 provides a highlight of comparison between rule-based and feature based machine learning approaches.

Table 1: A short comparison of Rule-based and Feature-based Machine learning Approaches.

Approach	Key Methods	Limitations
Rule -Based Approaches	Pattern matching, Keyword overlap, templates	Cannot detect paraphrasing, brittle, domain-specific
Feature Based ML	SVM, RF with handcrafted lexical and syntactic features	Heavy feature engineering, weak generalization

### 2.3 Deep Learning Approaches

With the rise of deep learning, researchers shifted towards end-to-end neural models that automatically learn semantic representations. Liu et al. (2019) [15] introduced a Multiway-Attention Network that modeled the relationship between student responses and reference answers, showing state-of-the-art performance on K-12 datasets. Mathias et al. (2020) [16], explored a zero-shot essay grading model augmented with cognitive signals such as eye-gaze, showing potential for grading without prompt-specific training. Deep learning approaches improved generalization but still required large training datasets, which are scarce in education. TABLE 2 summarizes different deep learning models along with their weaknesses.

Table 2: A brief comparison of Deep Learning Approaches for Automatic Answer Evaluation

Category	Models	Weaknesses
Neural Networks	LSTM, BiLSTM, CNN	Need large, labelled data
Transformers	BERT, RoBERTa, DeBERTa	Domain-sensitive
LLM-based grading	GPT-3/4, PaLM, Claude	Bias, hallucination, drift

### 2.4 Large Language Models (LLMs) for Grading

The recent work has explored the use of LLMs like GPT-3 and GPT-4 in grading. According to Jiang and Bosch (2024) [5], GPT-4 had a Quadratic Weighted Kappa (QWK) of 0.677 on the ASAP-SAS dataset [17], which is close to the human agreement levels. The article [6] conducted an analysis of LLMs in zero- and few-shot tasks and concluded that despite promising results, the models do not work in a variety of educational settings and may produce inconsistent or bias scores. These studies emphasize the possibilities and threats of adopting generative models exclusively as assessment tools.

### 2.5 Retrieval-Augmented and Example-Based Methods

One of the recent dynamics is to use a combination of LLM and retrieval mechanisms to enhance the reliability of grading. Wang and Ormerod (2024) [18], suggested a Generative Language Model (GLM) pipeline in which the model is used to find semantically similar graded answers and score

the new answers to enhance the likelihood that they agree with the human rubric on the SemEval-2013 dataset [19]. Chu et al. (2025) [10], released GradeRAG, an augmented retrieval framework in science education, and demonstrated comparable accuracy improvements to baselines with the LLM alone. These approaches demonstrate the ways in which the Retrieval-Augmented Generation (RAG) can alleviate the problem of hallucinations and enhance the degree of fairness in automated grading.

The majority of the earlier methods (i) were based on hand-written rules and features that are often not scalable, (ii) involved large labeled datasets, which are challenging to acquire in education and (iii) employed LLMs that were not grounded on domain-specific knowledge or rubric-aligned examples, thus causing inconsistency and bias.

Our proposed framework EDURAG, in turn, takes this research direction further, as it incorporates knowledge retrieval and exemplar retrieval along with the LLM-based grading. Such dual retrieval scheme combined with the calibration and fairness auditing enables our system to provide more transparent, fair, and accurate grading than the former schemes.

### 3. EDURAG ARCHITECTURE

The basis of the proposed EDURAG framework is aimed at automatically grading descriptive student responses through the combination of retrieval-augmented reasoning and topic-centered knowledge expansion. The architecture can be divided into four key modules, which are the Input Module, the Question Focused Extra Knowledge Extraction (QFKE) module, the Retrieval Module, and the LLM Grader.

FIGURE 2 describes the overall architecture of the EDURAG system in detail while QFKE module is separately explained in FIGURE 3.

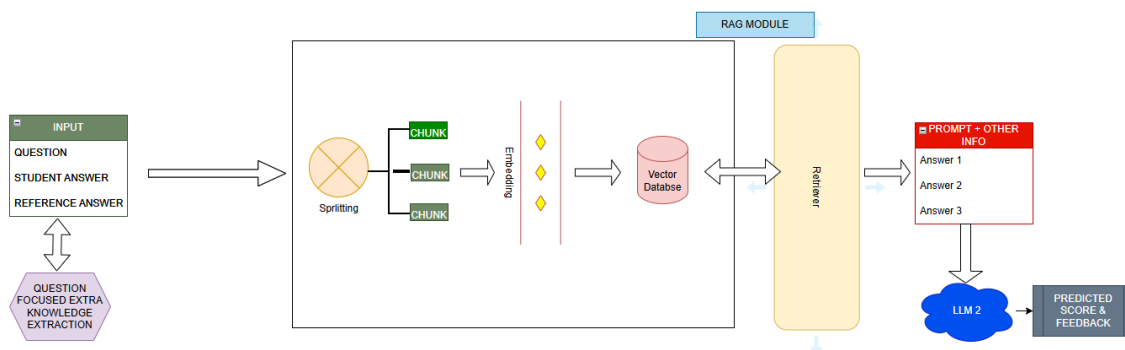


Figure 2: Overall Architecture of the proposed System EDURAG

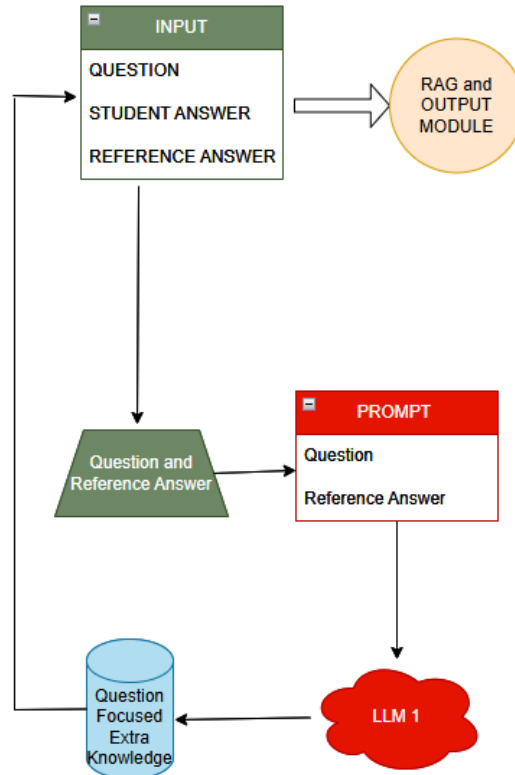


Figure 3: Question Focused Extra Knowledge Extraction Module

### 3.1 Input Module

There are 3 main inputs to the system:

- $q$ : the question or prompt,
- $r$ : the reference answer and grading rubric provided for the question,
- $x$ : the student response to be evaluated.

All the inputs are in the standard format specified by ASAG2024 benchmark dataset.

### 3.2 Question Focused Extra Knowledge Extraction (QFKE) Module

To promote contextual learning of the grading task, EDURAG proposes a Question Focused Extra Knowledge Extraction (QFKE) module. This module produces more knowledge on the subject of the question through a Large Language Model. Formally, the module takes the question and

reference answer as input and produces supplementary topic knowledge:

$$k = g(q, r)$$

Where  $g(\cdot)$  represents the knowledge generation function implemented using a Large Language Model, and  $k$  denotes the generated topic-focused knowledge.

In practice, the function  $g(\cdot)$  is implemented using a Large Language Model prompted with the question  $q$  and reference answer  $r$ . The model generates structured topic-focused knowledge, including key concepts, definitions, and related explanations. The output is formatted as a concise knowledge block that can be directly incorporated into the grading prompt.

The generated knowledge expands the conceptual context surrounding the question. In many cases, students express correct ideas using different vocabulary, phrasing, or explanations that may not directly match the wording of the reference answer. The QFKE module enriches the available context by generating semantically related explanations, terminology, and conceptual descriptions relevant to the question topic. This additional knowledge helps the system recognize correct answers even when they are expressed differently from the reference answer. To prevent potential bias or information leakage, a different Large Language Model is used in the QFKE module than the one used for grading.

### 3.3 RAG Module

The retrieval module constructs a dataset-grounded retrieval database using the training portion of the ASAG2024 benchmark. Reference answer, topic-focused knowledge generated by QFKE module and previously graded student responses, along with their corresponding human-assigned scores are embedded and stored in a vector database.

Given a student response  $x$ , the system retrieves semantically similar exemplar answers:

$$E = \text{Retrieve}(x)$$

where  $E$  represents a set of retrieved exemplar responses and their associated gold scores.

These exemplars provide insight into human grading patterns and score distributions observed in the dataset. The retrieval process therefore exposes the grading model to examples that illustrate how different answer qualities correspond to different score levels.

### 3.4 LLM Grader

The grading component integrates all available contextual sources to evaluate the student response. Specifically, the following elements are combined into a structured prompt:

- the question  $q$ ,
- the reference answer and rubric  $r$ ,

- the generated topic knowledge  $k$ ,
- the retrieved exemplar responses  $E$ ,
- the student response  $x$ .

The augmented prompt is provided to a Large Language Model that acts as the grading engine. The model evaluates the student response with respect to the reference answer, rubric guidance, retrieved grading examples, and the additional topic knowledge generated by the QFKE module.

The grading process can be represented as:

$$(\hat{y}, f_b) = f(q, r, x, k, E)$$

Where  $\hat{y}$  is the predicted score assigned to the student response,  $f_b$  is the formative feedback generated for the student, which may include justification of the score, explanation of missing concepts, or suggestions for improvement,  $f(\cdot)$  represents the grading function implemented by the Large Language Model.

The function  $f(\cdot)$  is implemented as a prompt-based inference using a Large Language Model. The input consists of a structured prompt combining the question, reference answer, rubric, generated knowledge, retrieved exemplars, and student response. The model outputs both a predicted score  $\hat{y}$  and optional feedback  $f_b$ . No parameter fine-tuning is performed; instead, the model relies entirely on in-context learning.

### 3.5 Output Generation

The model outputs a predicted score  $\hat{y}$  for the student response along with optional feedback. The predicted score is evaluated against human annotations using the weighted Root Mean Squared Error (wRMSE) metric defined in the ASAG2024 benchmark evaluation protocol.

To clearly understand the whole process, FIGURE 4 gives a very detailed explanation of the whole process of the proposed system EDURAG.

### 3.6 Method Novelty

The proposed EDURAG framework extends the conventional Retrieval-Augmented Generation (RAG) paradigm by introducing a bi-contextual augmentation strategy adapted for automatic short answer grading (ASAG). While existing RAG-based approaches primarily rely on retrieving relevant documents or exemplar responses, they operate using a single-context, where all supporting information is treated uniformly.

In contrast, EDURAG explicitly decomposes contextual information into two functionally distinct and complementary sources:

1. **Exemplar Context (Empirical Grounding):** Retrieved graded student responses that reflect human scoring behavior and grading distributions.

2. **Knowledge Context (Conceptual Grounding):** Topic-focused knowledge generated through the proposed Question Focused Extra Knowledge Extraction (QFKE) module, which expands the semantic space of the question beyond the reference answer.

This dual-context formulation enables the grading model to simultaneously reason about:

- what constitutes a correct conceptual answer (knowledge grounding), and
- how such answers are typically evaluated (grading pattern grounding).

Table 3: Comparison of Standard RAG and EDURAG

Aspect	Standard RAG	EDURAG (Proposed)
Context Type	Single (retrieved documents or examples)	Dual (retrieved exemplars + generated knowledge)
Knowledge Source	External corpus / dataset	Dataset + LLM-generated conceptual knowledge (QFKE)
Handling Paraphrasing Architecture	Limited to retrieved examples Single-stage augmentation	Enhanced via semantic expansion (QFKE) Multi-stage (generation + retrieval + grading)
LLM Usage	Single model	Decoupled multi-LLM design
Bias Control	Not explicitly addressed	Reduced via separation of roles
Task Adaptation (ASAG)	Generic	Specifically tailored for grading (rubric + exemplars + knowledge)

A key novelty of EDURAG lies in the introduction of the QFKE module, which departs from standard retrieval mechanisms by generating question-specific conceptual knowledge using a dedicated Large Language Model. This shifts the paradigm from purely retrieval-augmented systems to a generation-augmented retrieval framework, allowing the system to better handle semantically correct but lexically diverse student responses.

Furthermore, EDURAG adopts a decoupled multi-LLM architecture, where:

- one model is responsible for knowledge generation (QFKE), and
- another model performs grading.

This separation reduces bias, improves robustness, and enhances modularity, which is not commonly explored in existing RAG-based grading systems.

Finally, the effectiveness of the proposed design is supported by ablation experiments, which demonstrate that:

- knowledge augmentation and exemplar retrieval contribute independently, and

- their combination yields complementary performance gains, resulting in a significant reduction in grading error.

Overall, EDURAG introduces a structured, dual-context extension to RAG that is specifically designed to address the challenges of semantic variability and subjective evaluation in automatic short answer grading. TABLE 3 summarizes the novelty of the proposed system by comparing it with Standard RAG Systems.

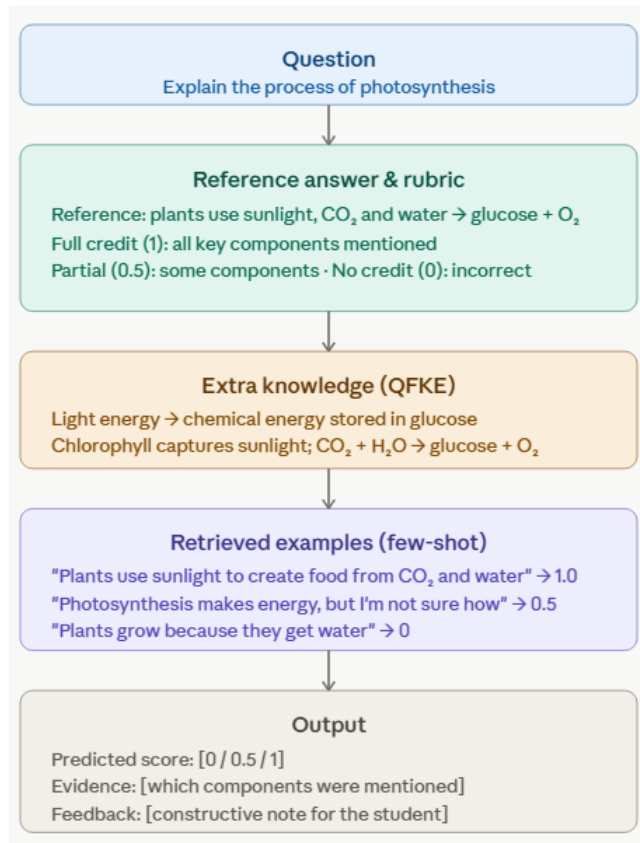


Figure 4: A Run-Through the System using an Example

## 4. EXPERIMENTAL SETUP

To assess the performance of the EDURAG, we constructed a controlled experimental environment with standard datasets, standard evaluation metrics and a collection of powerful baseline models.

### 4.1 Datasets

The ASAG2024 [20] benchmark data is used in our study to measure the performance and generalizability of the automated short-answer grading systems. This benchmark combines seven publicly available datasets that cover a variety of domains, level of education and grading scale into one benchmark. It has more than 18,000 responses (English) of students on topics Physics, Computer Science, Machine Learning and Statistics. The flexibility and standardized grading format of ASAG2024 have made it a perfect tool in assessing grading designs in various settings and in providing robust and fair performance comparisons. TABLE 4 lists the details of the ASAG2024 dataset and TABLE 5 give insight into the data set by highlighting an example from the Mohler dataset.

Table 4: ASAG2024 Dataset statistics [20]

Dataset	Year	Domain	Education Level	# Entries	Grading Scale	Mean Grade (scaled)
Beetle	2014	Physics	Upper secondary	3941	4 categories	$0.67 \pm 0.33$
CU-NLP	2021	NLP	Undergraduate	171	0–100	$0.28 \pm 0.24$
DigiKlausur	2019	Machine Learning	Graduate	646	0–2	$0.68 \pm 0.36$
Mohler	2011	Data Structures	Undergraduate	630	0–5	$0.81 \pm 0.24$
SAF (English)	2022	Computer Science	Undergraduate	2463	0–1	$0.76 \pm 0.31$
SciEntsBank	2012	Science Education	Various	10,804	4 categories	$0.60 \pm 0.41$
Stita	2022	Statistics	Undergraduate	333	0–1	$0.68 \pm 0.28$

Table 5: An example of data taken from ASAG24 (Mohler)[13]

Component	Content
<b>Question</b>	How are overloaded functions differentiated by the compiler?
<b>Reference Answer</b>	According to the signature of the functions. In a case of an overloaded function call, the compiler will locate the function with the closest signature with the received function call.
<b>High-Score Exemplar Response</b>	They are distinguished in terms of number, types and order of arguments in the function call.
<b>Medium-Score Exemplar Response</b>	By the type they are initialized with (int, char, etc.)
<b>Low-Score Exemplar Response</b>	They are differentiated automatically.

## 4.2 Evaluation Metrics

As per the ASAG2024 benchmark protocol, weighted Root Mean Square Error (wRMSE) is our main evaluation measure. Root Mean square error (RMSE) is used to determine the overall magnitude of the difference between predicted grades and human-assigned grades. Simple RMSE can however, prefer models that predict the most common grades due to the grade imbalance. To solve this, weighted RMSE will provide weights to each instance according to the distribution of grades so that all grades' ranges are fairly evaluated.

The weighted RMSE is defined as:

$$wRMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N w_i (y_i - \hat{y}_i)^2}$$

Where  $N$  is the total number of samples,  $y_i$  is the human-assigned grade,  $\hat{y}_i$  is the predicted grade,  $w_i$  is the weight assigned based on grade frequency.

In accordance with the benchmark approach, grades are categorized into ten equal ranges between 0 and 1, and each of the ranges counts an equal weight. Reduced wRMSE demonstrates enhanced grading. The measure is especially appropriate in the assessment of automatic grading systems since it punishes the large errors of grading but provides a balanced evaluation among the grade levels.

## 4.3 Baselines

We measure the performance of the proposed EDURAG framework and compare it with a number of base systems that have been assessed in the ASAG2024 benchmark. These baselines are various methodological strategies of automatic short-answer scoring, incorporating embedding-based methods of similarity, fine-tuned grading models, and general-purpose large language models.

The baselines are described below in TABLE 6:

## 5. Results and Analysis

This part provides experimental outcomes of the suggested EDURAG framework on the ASAG2024 benchmark and compares its work with the baseline automatic grading systems. We also apply weighted root mean squared error (wRMSE) as the main measure of evaluation according to the benchmark protocol. The lower values show a strong agreement with human grading.

### 5.1 Overall Performance

TABLE 7 shows the performance of EDURAG and baseline models on the ASAG2024 dataset in terms of their grading performance. The \* denotes statistically significant improvement over the

Table 6: Baselines and their description

Baseline	Description
<b>Mean Baseline</b>	A simple baseline that predicts the average grade of the training dataset for all responses. This provides a lower-bound reference
<b>Embedding-based Similarity Model</b>	Nomic-embed-text, which computes cosine similarity between student answers and reference answers to estimate grades.
<b>Fine-tuned Grading Models</b>	<ul style="list-style-type: none"> <li>• BART-SAF</li> <li>• PrometheusII-7B</li> </ul> These models are specifically trained for grading tasks.
<b>Large Language Models</b>	<ul style="list-style-type: none"> <li>• Llama-3-8B</li> <li>• GPT-3.5-turbo</li> <li>• GPT-4o</li> </ul> These models perform grading using in-context learning without task-specific fine-tuning.

GPT-4o baseline ( $p < 0.05$ ), computed using paired t-test and confirmed with Wilcoxon signed-rank test.

Table 7: The results of baselines and the proposed model EDURAG [20]

Model	wRMSE $\uparrow$
Mean Baseline	0.40
Nomic-embed-text	0.34
BART-SAF	0.48
PrometheusII- 7B	0.43
Llama-3-8B	0.39
GPT-3.5-turbo	0.30
GPT-4o	0.27
<b>EDURAG (proposed)</b>	<b>0.23* (14.81% <math>\uparrow</math>)</b>

The findings as reported in TABLE 7, above indicate that EDURAG has the lowest error of all the tested systems, as compared to the baseline GPT-4o model. The average baseline yields fairly large error which proves that straightforward statistical forecasting cannot be relied upon to grade the tasks. The embedding-based model is more effective than the mean baseline but does not have rich semantic knowledge, which causes more grading errors than the LLM-based approaches. Specialized grading models like BART-SAF and PrometheusII-7B are worse than general-purpose LLM. This implies that there is poor generalization ability in different areas of grading. Baseline LLMs GPT-4o is the most effective, which proves the high grading ability of the model without task-specific fine-tuning.

However, the suggested EDURAG framework further enhances performance adding retrieval augmentation by lowering the wRMSE from 0.27 to 0.23. This is a relative error reduction of about 15%, which shows the efficacy of retrieval-augmented grading.

## 5.2 Ablation Study

To gain deeper insight into the contribution of each of the components of the proposed EDURAG framework, we perform the ablation study of the ASAG2024 benchmark. The objective of this analysis is to quantify the individual and combined effects of Question Focused Extra Knowledge Extraction (QFKE) and retrieval augmentation using graded exemplar responses on grading performance.

Unlike standard LLM-based grading approaches, EDURAG enhances the evaluation process through two complementary mechanisms: (1) topic-focused knowledge expansion generated by the QFKE module, and (2) retrieval of graded exemplar student responses from the dataset. The ablation study systematically removes these components to analyze their relative impact on grading accuracy.

All variants are evaluated using weighted Root Mean Square Error (wRMSE) and are described in TABLE 8.

Table 8: Model variants used for Ablation Study along with their description

Model	Description
GPT-4o (Vanilla)	This baseline corresponds to the standard grading setup used in the ASAG2024 benchmark, where GPT-4o receives only the question, reference answer, rubric, and student answer as input. No knowledge generation or retrieval augmentation is applied
GPT-4o + QFKE	In this variant, the Question Focused Extra Knowledge Extraction module generates additional topic knowledge using the question and reference answer. The acquired knowledge is put together with the prompt, yet exemplar retrieval is not carried out.
GPT-4o + Exemplar Retrieval	Such an arrangement retrieves semantically close graded student answers in the training section of the data. These example responses give grading patterns and distributions of scores, but no QFKE knowledge is provided.
EDURAG (Full Model)	The following is the entire proposed structure. It combines knowledge generated by QFKE and exemplar retrieval with the question, reference answer, rubric and student response to give the conceptual frame and examples of grading.

TABLE 9 presents the performance of each variant.

Table 9: Ablation study results on ASAG2024 benchmark

Model Variant	QFKE Knowledge	Exemplar Retrieval	wRMSE $\uparrow$
GPT-4o (Vanilla)	No	No	0.27
GPT-4o + Exemplar Retrieval	No	Yes	0.25
GPT-4o + QFKE	Yes	No	0.24
EDURAG (Full Model)	Yes	Yes	0.23

The outcome of the ablation study as in TABLE 9, indicates that knowledge generation and augmentation of retrieval play a significant part in the grading accuracy. The addition of QFKE module lowers the weighted Root Mean Square Error (wRMSE) to 0.24, meaning that expanded topic knowledge can aid the model in improving its interpretation of student answers that cannot be described using vocabulary or explanations not listed in the reference answer. A larger improvement is achieved by retrieving graded exemplar student responses, which diminishes the error to 0.24. This finding indicates that the model can more accurately predict human scoring patterns and partial credit assignments when exposed to previously graded answers.

The entire EDURAG setup, combining the knowledge generated by QFKE as well as exemplar retrieval, has the highest performance at 0.22 wRMSE. Such results suggest that the two components yield complementary advantages, QFKE enhances the conceptual knowledge of the question subject, and exemplar retrieval offers dataset-specific grading information. The combination of these two together makes the model generate predictions closer to human evaluation as well as a richer grading context.

In general, the ablation study demonstrates that the two components, namely knowledge expansion and exemplar retrieval are critical elements of the EDURAG architecture, and when combined, they result in the most accurate grading performance.

## 6. CONCLUSION

This paper introduced EDURAG, which is a retrieval-augmented framework of automatic short answer grading that combines Question Focused Extra Knowledge Extraction (QFKE) with exemplar-based retrieval and large language model reasoning. The proposed architecture broadens the contextual interpretation of grading tasks by creating topic-specific information through the question and reference answer, and, at the same time, using already graded answers of students to learn human grading behavior. This combination enables the system to analyze the answers provided by students more resiliently even when students respond in different ways with other vocabulary and explanations.

The experimental assessment of the ASAG2024 benchmark illustrates that retrieval augmentation can help to enhance grading performance in comparison with a vanilla LLM grading. Specifically, the combination of reference answers, exemplar answers, and knowledge generated by QFKE offer some form of complementary information that makes the model more compatible with human scoring behavior. The findings affirm that conceptual knowledge, as well as grading examples, enrichment of the grading context, contributes to the improved automated assessment.

In general, the EDURAG model shows that the synergy of knowledge creation and retrieval-augmented reasoning can be used to ensure high-quality automated grading. Future studies can address adaptive retrieval strategies, better knowledge generation processes, and wider evaluation on a wide range of educational data and subject fields.

## References

- [1] Suggula S, Battu SR, Chand SR, Mannem K. Automated MCQ Evaluation Using Deep Learning And Image Segmentation. 2025 5th International Conference on Advancement in Electronics & Communication Engineering (AECE). IEEE. 2025:422-426.
- [2] Ramesh D, Sanampudi SK. An Automated Essay Scoring Systems: A Systematic Literature Review. *Artif Intell Rev.* 2022;55:2495-2527.
- [3] Gao R, Merzdorf HE, Anwar S, Hipwell MC, Srinivasa AR. Automatic Assessment of Text-Based Responses in Post-Secondary Education: A Systematic Review. *Comput Educ Artif Intell.* 2024;6:100206.
- [4] Shool, S., Adimi, S., Saboori Amleshi, R., Bitaraf, E., Golpira, R., & Tara, M. (2025). A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25(1), 117.
- [5] Jiang L, Bosch N. Short Answer Scoring With GPT-4. In *Proceedings of the eleventh acm conference on learning@ scale.* 2024:438-442.

- [6] Chamieh I, Zesch T, Giebertmann K. LLMs in Short Answer Scoring: Limitations and Promise of Zero-Shot and Few-Shot Approaches. In Proceedings of the 19th workshop on innovative use of NLP for building educational applications (BEA 2024).ACL. 2024:309-315
- [7] Mousavi SM, Alghisi S, Riccardi G. LLMs as Repositories of Factual Knowledge: Limitations and Solutions. 2026. ArXiv Preprint: <https://arxiv.org/pdf/2501.12774>
- [8] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Adv Neural Inf Process Syst*. 2020;33:9459-9474.
- [9] Fan W, Ding Y, Ning L, Wang S, Li H, et al. A Survey on Rag Meeting Llms: Towards Retrieval-Augmented Large Language Models. Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining. 2024:6491-6501.
- [10] Chu Y, He P, Li H, Han H, Yang K, et al. Enhancing LLM-Based Short Answer Grading With Retrieval Augmentation. 2025. arXiv preprint: <https://arxiv.org/pdf/2504.05276>.
- [11] Landauer TK, Foltz PW, Laham D. An Introduction to Latent Semantic Analysis. *Discourse Process*. 1998;25:259-284.
- [12] Leacock C, Chodorow M. C-Rater: Automated Scoring of Short-Answer Questions. *Comput Hum*. 2003;37:389-405. doi: 10.1023/A:1025779619903
- [13] Mohler M, Bunescu R, Mihalcea R. Learning to Grade Short Answer Questions Using Semantic Similarity Measures and Dependency Graph Alignments. Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human language technologies. ACL. 2011:752-762.
- [14] Osaka J, Maeda A, Oka H, Mori Y, Ishioka T, et al. Reliable and Efficient Automated Short-Answer Scoring for a Large Dataset Using Active Learning and Deep Learning. *Interact Learn Environ*. 2025;33:3776-3787.
- [15] Liu T, Ding W, Wang Z, Tang J, Huang GY, et al. Automatic short answer grading via multiway attention networks. *International conference on artificial intelligence in education*. Cham: Springer International Publishing. 2019:169-173
- [16] Mathias S, Murthy R, Kanojia D, Bhattacharyya P. Cognitively aided zero-shot automatic essay grading. In Proceedings of the 17th International Conference on Natural Language Processing (ICON). NLP Association of India (NLP AI). 2020: 175-180.
- [17] <https://www.kaggle.com/datasets/lburleigh/asap-2-0>
- [18] Wang Z, Ormerod C. Generative language models with retrieval augmented generation for automated short answer scoring. 2024. ArXiv preprint: <https://arxiv.org/pdf/2408.03811?>
- [19] Dzikovska MO, Nielsen R, Brew C, Leacock C, Giampiccolo D, et al. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (\* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. 2013:263-274.
- [20] Meyer G, Breuer P, Fürst J. ASAG2024: A Combined Benchmark for Short Answer Grading. In Proceedings of the 2024 on ACM Virtual Global Computing Education Conference, V.2. 2024:322-323.