Enhancing LLM-Based Short Answer Grading with Retrieval-Augmented Generation

Yucheng Chu
Michigan State University
chuyuch2@msu.edu

Peng He
Washington State University
peng.he@wsu.edu

Hang Li
Michigan State University
lihang4@msu.edu

Haoyu Han
Michigan State University
lhanhaoy1@msu.edu

Kaiqi Yang
Michigan State University
kqyang@msu.edu

Yu Xue
Washington State University
yu.xue@wsu.edu

Tingting Li
Washington State University
tingting.li1@wsu.edu

Yasemin Copur-Gencturk
University of Southern California
copurgen@usc.edu

Joseph Krajcik
Michigan State University
krajcik@msu.edu

Jiliang Tang
Michigan State University
tangjili@msu.edu

ABSTRACT

Short answer assessment is a vital component of science education, allowing evaluation of students' complex three-dimensional understanding. Large language models (LLMs) that possess human-like ability in linguistic tasks are increasingly popular in assisting human graders to reduce their workload. However, LLMs' limitations in domain knowledge restrict their understanding in task-specific requirements and hinder their ability to achieve satisfactory performance. Retrieval-augmented generation (RAG) emerges as a promising solution by enabling LLMs to access relevant domain-specific knowledge during assessment. In this work, we propose an adaptive RAG framework for automated grading that dynamically retrieves and incorporates domain-specific knowledge based on the question and student answer context. Our approach combines semantic search and curated educational sources to retrieve valuable reference materials. Experimental results in a science education dataset demonstrate that our system achieves an improvement in grading accuracy compared to baseline LLM approaches. The findings suggest that RAG-enhanced grading systems can serve as reliable support with efficient performance gains.

Keywords

Automated Short Answer Grading, LLM, RAG, Learning Assessments, Constructed Responses

1. INTRODUCTION

Assessment and analysis of student understanding in science education extend far beyond simple grading. With the emergence of new frameworks for K-12 science education, there is an increasing emphasis on analyzing students' multidimensional comprehension of scientific understanding. The National Research Council's Framework for K-12 Science Education [4] established three critical dimensions of science learning: disciplinary core ideas (DCIs), science and engineering practices (SEPs), and crosscutting concepts (CCCs). To demonstrate their achievement, students should apply their 3D understanding to interpret compelling phenomena or solve real-world problems. This framework has fundamentally transformed how we assess student learning, shifting away from traditional multiple-choice questions toward short-answer assessments that better capture students' authentic understanding [8]. Short-answer questions have become particularly valuable in this context, as they require students to demonstrate their three-dimensional knowledge by explaining real-world phenomena and solving complex problems [10]. These responses provide rich insights into students' conceptual understanding, revealing both their mastery and misunderstanding, which is essential for teachers to adapt their instructional strategies and support students' self-regulated learning [9].

However, the analysis of these short answers presents significant challenges as it is time-intensive, requires deep expertise across all three dimensions, and makes it difficult for teachers to provide timely, actionable feedback to individual students. Fortunately, artificial intelligence (AI), particularly automatic short answer grading (ASAG) systems, has emerged as a promising solution to these challenges. Traditional ASAG approaches using machine learning (ML) techniques have shown success in providing consistent and objective scoring but are limited by their reliance on large training datasets [2, 19]. The recent advent of large language models (LLMs) has opened new possibili-

ties for training ASAG systems on smaller datasets, due to its possession of broad general knowledge from pre-training. However, current LLM-based approaches face two critical limitations in science education assessment. First, they exhibit deceptive performance in technical domains, often generating plausible-sounding but scientifically false evaluations when in-depth domain knowledge is required. Their output can be unreliable when evaluating answers that contain ambiguous terminology or complex scientific concepts. Second, they lack sufficient task-specific knowledge about the grading criteria [13]. They often fail to understand the nuanced requirements of specific educational frameworks such as the three-dimensional learning approach, which can lead to misalignment between automated scores and expert evaluations.

To address these limitations, we introduce GradeRAG, a novel framework that enhances LLM-based grading through retrieval-augmented generation (RAG). Our approach implements a specialized RAG pipeline that provides the LLM with access to curated domain-specific knowledge bases, enabling a more accurate assessment of scientific concepts across the three dimensions (DCIs, SEPs, and CCCs). This knowledge-focused approach is particularly crucial in science education, where effective assessment requires domain-specific expertise of specialized concepts and their interconnections [10, 1]. As part of our knowledge retrieval strategy, we incorporate expert-annotated scoring rationales as a specialized form of knowledge source. These examples contain detailed scoring rationales that explicitly identify critical scientific terminology and reasoning patterns in student responses. By treating these examples as retrievable knowledge, LLMs are guided to emulate expert analysis processes, further improving scoring performance. We evaluate GradeRAG on a dataset of student short answers on science assessments. Results demonstrate improvements in both grading accuracy and consistency compared to the baseline approaches. Our results suggest that integrating specialized knowledge retrieval systems can bridge the gap between automated efficiency and expert-level assessment in science education, leading to more reliable evaluation of complex scientific understanding.

2. RELATED WORK

Automatic short answer grading. Automatic short answer grading (ASAG) has evolved significantly over the past decades. Early approaches mainly focus on text matching and statistics-based methods [17]. With the advances in ML, later systems employ feature engineering techniques with supervised training methods [11]. These traditional approaches typically require large training datasets and often struggle with capturing the semantic differences in student answers. Recent works utilizing neural networks demonstrate improvement in understanding semantic relationships in short answers [7]. Large language models (LLMs) further enable enhanced capabilities in ASAG through zero-shot and few-shot [12] techniques. However, as noted by [15, 14], LLMs still face challenges in specialized domains like science education, where domain-specific knowledge is crucial for accurate assessment. Recent work introduces GradeOpt, a multi-agent LLM framework that iteratively refines grading guidelines to incorporate expert knowledge and achieve higher alignment with human graders [3]. Complementary

research leverages multi-agent LLM collaboration to identify domain-specific content knowledge gaps, demonstrating the value of agent-based methods for improving assessment quality [21].

Retrieval-augmented generation. Retrieval-augmented generation (RAG) has emerged as an effective approach to enhance LLMs' reasoning ability by providing access to external knowledge sources. RAG has demonstrated effectiveness in knowledge-intensive tasks requiring factual accuracy [13] and domain expertise [5]. In the educational context, the application of RAG remains relatively unexplored, with most works focusing on content generation [16] rather than assessment. Recent studies have begun investigating RAG for automated assessment. For example, [6] proposes a rubric-centric approach for automated test correction using RAG, while [20] evaluated RAG frameworks for grading open-ended written responses. However, the application of RAG, specifically for short answer grading in science education, where domain-specific knowledge across multiple dimensions (DCIs, SEPs, CCCs) is crucial, represents a novel contribution to our work.

3. METHOD
3.1 Problem Statement
The task of automatic short answer grading (ASAG) involves assigning appropriate scores to open-ended student responses. Given a student answer x, an ASAG system maps it to one of C predefined score levels $\{s_1, ..., s_C\}$. Formally, an ASAG system $\mathcal{F}$ generates score predictions $\hat{y}_i = \mathcal{F}(x_i)$ for each response x_i across each dimension. This work extends traditional ASAG through retrieval-augmented generation. We leverage two types of external knowledge resources to enhance grading performance. First, we utilize domain-specific knowledge from an external database K_D, retrieving relevant background information I that contains educational materials and scoring guidelines. Second, we incorporate a specialized knowledge collection K_E of N expert-annotated graded examples $\mathcal{E} = \{(x_i, r_i, y_i)|i = 1, ..., N\}$, where each entry consists of a student response x_i, a scoring rationale r_i, and the corresponding ground truth score y_i. By combining these knowledge sources, our enhanced ASAG system $\mathcal{F}(x_i, I, \mathcal{E})$ creates a more comprehensive context for LLM-as-a-Grader's accurate assessment.

3.2 RAG-Based Grading System
In this section, we present GradeRAG, a framework that combines in-context learning with RAG for automatic short answer grading in science education.

3.2.1 Knowledge Base Construction
As shown in Figure 1, our system constructs a triple-index knowledge base optimized for scientific assessment. K_{D_1} and K_{D_2} consists of project-related information and grading guidelines, while K_E (which is optinal) contains expert-annotated scoring examples. The first index K_{D_1} contains general assessment documents such as the Framework of K-12 Science Education [4] and the NGSS standard document [18]. These materials are systematically chunked into coherent segments of approximately 512 tokens, which balances content completeness with the retrieval granularity. This chunking size preserves sufficient context while
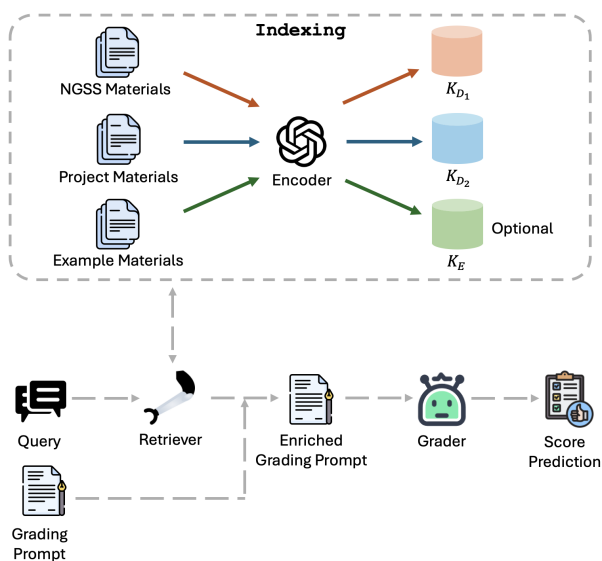
Figure 1: An illustration of the proposed GradeRAG framework.

Figure 2: An example prompt for grader

enabling precise retrieval of relevant concepts, as smaller chunks would fragment scientific explanations while larger chunks can introduce excessive irrelevant information. Each chunk is tagged with metadata including the document source and dimension type to facilitate efficient filtering during retrieval.

The second index K_{D_2} consists of task-specific materials, such as the three-dimensional learning progression materials [10], which includes unpacking DCIs, SEPs, and CCCs materials and the holistic levels in the three assessment tasks. For these more specialized resources, we employ a finer-grained, manual chunking strategy. Specifically, we divide the content into single sentences or meaningful segments containing one or two bullet points. This approach is adopted as task-specific materials contain dense, concentrated information where even small segments can provide critical assessment criteria. With this fine-grained chunking strategy, our system can retrieve precisely the relevant assessment standards without including extraneous information. Each chunk is tagged with metadata indicating the dimension type, level number, and chunk identifier to enable targeted filtering during retrieval.

The third index K_E comprises expert-annotated examples, each encoded with metadata such as score assignments and dimension-specific annotations. These examples are maintained as complete units rather than being chunked, which preserves the complete reasoning logic of expert scoring. This design choice ensures that our framework can access the full scoring demonstration with coherent guidance for emulating human-like reasoning processes, thus improving the alignment between model scores and human evaluation.

3.2.2 Dual Retrieval Strategy

For each student response x, GradeRAG employs a dual retrieval mechanism to gather complementary information.

Knowledge Retrieval. The system first identifies relevant dimension-specific content using a task-aware retrieval strategy. Each task is associated with specific dimension levels (e.g., Task 1 corresponds to DCI Level 1 and 2, while Task 2 encompasses Level 4, 5 and 6). This retrieval mechanism operates in two steps. First, it retrieves dimension-specific materials using a semantic similarity search over the indexed content that corresponds to different levels. The top-k pertinent information is retrieved. Second, to optimize relevance to the student answer while accounting for the context window constraint, we implement a reranking mechanism that combines three weighted components: semantic similarity (40% weight), text matching based on word overlap (30% weight), and domain-specific concept coverage (30% weight). We deliberately combine these three scores to address the different aspects of information relevance in science education assessment. Semantic similarity captures contextual relevance and identifies materials that share similar meanings. Text matching based on normalized word overlap between the student response and knowledge chunks can identify the lexical connections. This ensures that specific terminology in student answers is represented in the retrieved content. Lastly, the concept matching score evaluates the coverage of key scientific concepts necessary for assessment, such as properties, substances, chemical reactions, and identifications in the knowledge chunk. The final top-k passages are selected based on this combined score. By weighting and combining these three signals, our system achieves more comprehensive retrieval.

Example Retrieval. In parallel with knowledge retrieval, our system retrieves similar expert-annotated examples using semantic similarity matching. These examples include expert rationales that explain the scoring criteria and high-

light the critical keywords in student responses that signal understanding. Figure 3 shows a concrete example of the expert-annotated grading sample. We incorporate demonstrations (human-graded examples) into the grading prompt using a scale-based approach. The number of demonstrations scales with the complexity of the scoring scheme. For binary classification tasks with two score classes, we utilize six demonstrations, while for ternary classification tasks with three score classes nine demonstrations are employed. This scaling ensures sufficient examples across all possible score categories while maintaining prompt efficiency.

Examplar of Expert-Annotated Sample with Scoring Rationale (SEP)

Student's Short Answer: When coconut oil is mixed with lye, a chemical reaction occurs because soap and glycerol are new substances. From the table, I found that the odor, density, solubility in water, and melting point are different from each other. They are properties that can be used to identify substances and whether a chemical reaction occurs.

Score: SEP-1

Scoring Rationale: "When coconut oil is mixed with lye, a chemical reaction occurs because soap and glycerol are new substances." – this part meets the *partial* SEP criteria that mentioned a descriptive explanation, including a claim of a chemical reaction occurs and observed evidence of the data before and after the process. However, the response *did not use the evidence* to connect to the phenomenon – new substance produced and a chemical reaction occurred.

Figure 3: An example of expert-elaborated rationales for a short-answer sample on the SEP dimension.

3.2.3 Grading

After retrieving relevant knowledge and examples, GradeRAG performs the grading process. For each student response, GradeRAG assembles a unified grading context by combining the original expert-designed task guidelines, the retrieved top-k knowledge chunks, and similar expert-annotated examples with their rationales. This comprehensive content combination augments the expert-designed base grading guideline, providing additional context and criteria for more accurate assessment. The combined prompt (shown in Figure 2) is passed to the LLM-powered Grader agent for score generation. Formally, the Grader generates $\hat{y}_{1:T} = \mathcal{F}(x, K, E)$ where $\hat{y}_{1:T-1}$ represents the reasoning rationale containing $T-1$ tokens that explain the grading decision, and $\hat{y}_T$ is the final token that represents the numerical score. The input consists of the student answer x, retrieved knowledge chunks K, and expert examples E. The scoring process is performed separately for each dimension to ensure a focused evaluation of the different aspects of scientific understanding. The Grader's output includes both a numerical score $\hat{y}_T$ indicating the student's achievement level and a

detailed explanation $\hat{y}_{1:T-1}$ that references the specific evidence from the student response and connects it to the retrieved knowledge. This approach simulates the expert annotation process to enhance grading alignment.

4. EXPERIMENT

This section presents our experimental evaluation for our GradeRAG system. Specifically, we intend to investigate whether incorporating external knowledge through RAG improves automated scoring performance compared to the baseline approaches. We further investigate whether GradeRAG is compatible with *in-context learning* by conducting an ablation study.

4.1 Dataset

Our evaluation uses a testing dataset $\mathcal{D}_{test}$ comprising student responses collected from two middle schools in the Midwestern United States. The dataset includes student responses for each of the three NGSS-aligned science tasks, completed in standard classroom settings. The detailed response count for $\mathcal{D}_{test}$ is reported in Table 1. To establish ground truth scores, two content experts independently graded all responses using standardized rubrics and demonstration examples. Any scoring discrepancies were resolved through expert discussion to ensure high-quality gold standard annotations. Table 1 shows the details of level coverage for each question task.

Table 1: Detailed statistics of different questions in dataset $\mathcal{D}_{test}$. The number of samples in each label category is shown as C_i.

Question	Assessment Level		Response Count	
			Total	C1 / C2 / C3
Q_1	DCI	$(0, 1, 2)$		28/14/5
	SEP	$(0, 1, 2)$	47	44/3/0
	CCC	$(0, 1, 2)$		44/3/0
Q_2	DCI	$(0, 4, 5, 6)$		13/15/1/2
	SEP	$(0, 1, 2)$	31	12/3/16
	CCC	$(0, 1, 2)$		10/18/3
Q_3	DCI	$(0, 6, 7)$		40/3/3
	SEP	$(0, 1, 2, 3)$	46	12/26/5/3
	CCC	$(0, 1, 2, 3)$		14/1/28/3

4.2 Experimental Setting

We implement our grading system using gpt-4o-mini-2024-07-18 as the core Grader agent. To ensure reproducibility and consistent evaluation, we set the temperature parameter to 0, eliminating stochastic variations in the model's output. The embedding model used for indexing documents is text-embedding-ada-002. Since the original student answers were collected in image format, we first process them using gpt-4-vision-preview for transcription. These transcribed responses are then passed to the Retriever and Grader agent for separately assessing each task dimension (e.g., DCI, SEP, CCC). We run experiments for testing our system under two experimental conditions: NonRAG and GradeRAG. Under naive prompt, the system processes student responses using only the basic grading rubric without additional context or examples from the RAG component. Under GradeRAG

with zero-shot prompting, the system incorporates the retrieved materials as described in Section 3.2, but without expert-annotated examples.

Table 2: Performance comparison between NonRAG and GradeRAG on $\mathcal{D}_{test}$ with zero-shot prompting (shot number C=0, retrieval size k=4).

Question	NonRAG			GradeRAG		
	DCI	SEP	CCC	DCI	SEP	CCC
Accuracy (Acc)						
Q_1	0.217	0.717	0.739	0.348	0.804	0.957
Q_2	0.355	0.387	0.581	0.484	0.419	0.645
Q_3	0.717	0.478	0.565	0.783	0.522	0.674
Weighted F1 Score (F1)						
Q_1	0.170	0.799	0.816	0.257	0.858	0.935
Q_2	0.399	0.463	0.544	0.527	0.499	0.595
Q_3	0.752	0.443	0.543	0.792	0.481	0.656
Cohen's Kappa (κ)						
Q_1	-0.001	0.188	0.074	0.100	0.119	0.000
Q_2	0.129	0.232	0.188	0.245	0.260	0.288
Q_3	0.124	0.083	0.123	0.196	0.174	0.331

We evaluate performance using standard metrics, including accuracy, weighted F1-score, and Cohen's kappa. To be specific, Accuracy (Acc) measures the proportion of correct predictions. Weighted F1-score ($F1_{weighted}$) accounts for class imbalance by computing the F1-score for each class and taking their weighted average. Cohen's kappa (κ) measures the inter-rater reliability by accounting for the possibility of agreement occurring by chance. The formulas for these metrics are as follows:

$$Acc = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}_{(y_i=\hat{y}_i)}, \ \kappa = \frac{p_o - p_e}{1 - p_e},$$

$$F1_{weighted} = \sum k = 1^K n_k \cdot \frac{2 \cdot P_k \cdot R_k}{P_k + R_k}$$

where p_o is the observed agreement between the model and expert scores, p_e is the expected agreement by chance, n_k is the proportion of samples in class k, P_k and R_k are the precision and recall for class k.

4.3 Main Results

Table 2 presents the comparative performance of our proposed GradeRAG framework against the non-retrieval baseline (NonRAG) across three questions (Q_1, Q_2, and Q_3) and three assessment dimensions (DCI, SEP, and CCC). The results demonstrate general performance improvements with GradeRAG across metrics, questions, and dimensions. For all three metrics (accuracy, weighted F1 score, and Cohen's Kappa), GradeRAG outperforms the non-retrieval baseline in most cases. The most substantial gains are observed in Q_1 for the DCI dimension (13.1% accuracy increase) and CCC dimension (21.8% accuracy increase), with Q_3's CCC dimension showing the largest Kappa improvement (20.8% increase). While the Kappa values remain relatively low overall (ranging from 0.000 to 0.331), GradeRAG generally

achieves better alignment with expert grading patterns compared to NonRAG.

Analyzing performance across questions reveals patterns related to question complexity. Q_1, which has a relatively simple assessment structure with three levels (0, 1, 2) across all dimensions and highly imbalanced class distributions, demonstrates the largest average performance gains (14.5% in accuracy). This suggests that for questions with clear-cut scoring criteria and imbalanced response distributions, retrieved knowledge can provide particularly effective guidance. In contrast, Q_2 and Q_3 involve more complex assessment criteria with up to four score levels and more balanced class distributions, resulting in more moderate performance improvements. In general, this variety in complexity levels across questions suggests how retrieval effectiveness may depend on both the nature of the question and the distribution of student responses across performance levels.

Table 3: Performance comparison between NonRAG and GradeRAG under in-context learning (ICL) on $\mathcal{D}_{test}$.

Question	NonRAG-ICL			GradeRAG-ICL		
	DCI	SEP	CCC	DCI	SEP	CCC
shot number $C = 3$, retrieval size $k = 1$						
Accuracy (Acc)						
Q_1	0.609	0.543	0.826	0.609	0.565	0.913
Q_2	0.742	0.226	0.387	0.742	0.323	0.548
Q_3	0.674	0.500	0.304	0.696	0.587	0.348
Weighted F1 Score (F1)						
Q_1	0.591	0.664	0.875	0.599	0.682	0.927
Q_2	0.717	0.229	0.315	0.742	0.340	0.537
Q_3	0.726	0.473	0.300	0.741	0.555	0.327
Cohen's Kappa (κ)						
Q_1	0.230	0.087	0.281	0.303	0.094	0.292
Q_2	0.536	0.055	0.394	0.554	0.187	0.259
Q_3	0.170	0.108	0.142	0.189	0.231	0.142
shot number $C = 6$, retrieval size $k = 2$						
Accuracy (Acc)						
Q_1	0.609	0.696	0.891	0.630	0.717	0.913
Q_2	0.677	0.290	0.419	0.774	0.355	0.419
Q_3	0.761	0.413	0.391	0.761	0.413	0.348
Weighted F1 Score (F1)						
Q_1	0.584	0.786	0.913	0.609	0.798	0.927
Q_2	0.665	0.320	0.333	0.758	0.395	0.333
Q_3	0.785	0.400	0.317	0.784	0.392	0.272
Cohen's Kappa (κ)						
Q_1	0.210	0.053	0.238	0.276	0.178	0.292
Q_2	0.434	0.139	0.120	0.600	0.213	0.120
Q_3	0.259	-0.011	0.129	0.259	-0.006	0.068

Across dimensions, we observe that CCC assessments generally benefit most from knowledge augmentation, with an average improvement of 13.0% in accuracy. This aligns with expectations as crosscutting concepts often require broader

contextual understanding that can be enhanced through retrieved knowledge. The DCI dimension shows the second-largest improvements, while SEP shows more modest gains, suggesting that scientific practices may be more challenging to evaluate through retrieved content alone.

4.4 Ablation Study

In this section, we conduct two ablation experiments. First, to investigate whether the proposed GradeRAG is compatible with in-context learning (ICL), we conducted ablation experiments comparing NonRAG-ICL and GradeRAG-ICL across different shot settings. Tables 3 presents our findings with 3-shot and 6-shot settings.

Effect of Retrieved Knowledge Across Shot Settings. Comparing across shot settings reveals several patterns. In the zero-shot scenario, adding retrieved knowledge consistently improves performance across nearly all dimensions and questions. For 3-shot learning, the addition of knowledge retrieval also provides benefits, with GradeRAG-ICL outperforming NonRAG-ICL in 7 out of 9 dimension-question combinations. For DCI in both Q_1 and Q_2, retrieval provides no additional benefit, suggesting the examples may already contain sufficient domain knowledge. With 6-shot learning, GradeRAG-ICL outperforms NonRAG-ICL in only 5 out of 9 cases. For Q_3, knowledge retrieval provides no accuracy improvement in any dimension, even with a minor drop in CCC. This suggests that with sufficient examples, additional knowledge becomes redundant for this particular question.

Effect of Increasing Example Count. For NonRAG, increasing examples from 0 to 3 shots substantially improves performance in most cases, particularly for DCI dimensions. However, further increasing from 3 to 6 shots yields miscellaneous results, as we observe some dimensions improve while others deteriorate. For GradeRAG, the impact of increasing examples follows a similar pattern. The transitions from zero-shot to 3-shot generally improve performance, yet the magnitude of improvement is typically smaller than for NonRAG. This suggests that retrieval partially compensates for the lack of examples. The transitions from 3 to 6 shots within GradeRAG show varying effects. Positive gains appear in all questions in DCI and 2 out of 3 in SEP, yet performance decreases for 1 out of 3 in both SEP and CCC.

These two observations suggest the following findings regarding the compatibility of GradeRAG with in-context learning. First, adding more examples might lead to potential information excessiveness, causing diminishing utility or even performance decreases. Second, the optimal balance between retrieval and examples depends heavily on the specific dimension and question complexity. Third, simply adding more examples or more retrieved content does not guarantee improved performance. Different dimensions require different balances between the retrieval size and example number to achieve optimal performance.

Effect of Knowledge Retrieval Size. We further conducted a systematic ablation study on the impact of retrieval size k (number of retrieved knowledge chunks) across different in-context learning settings. Figures 4 represent the accuracy results for all questions across all dimensions as we
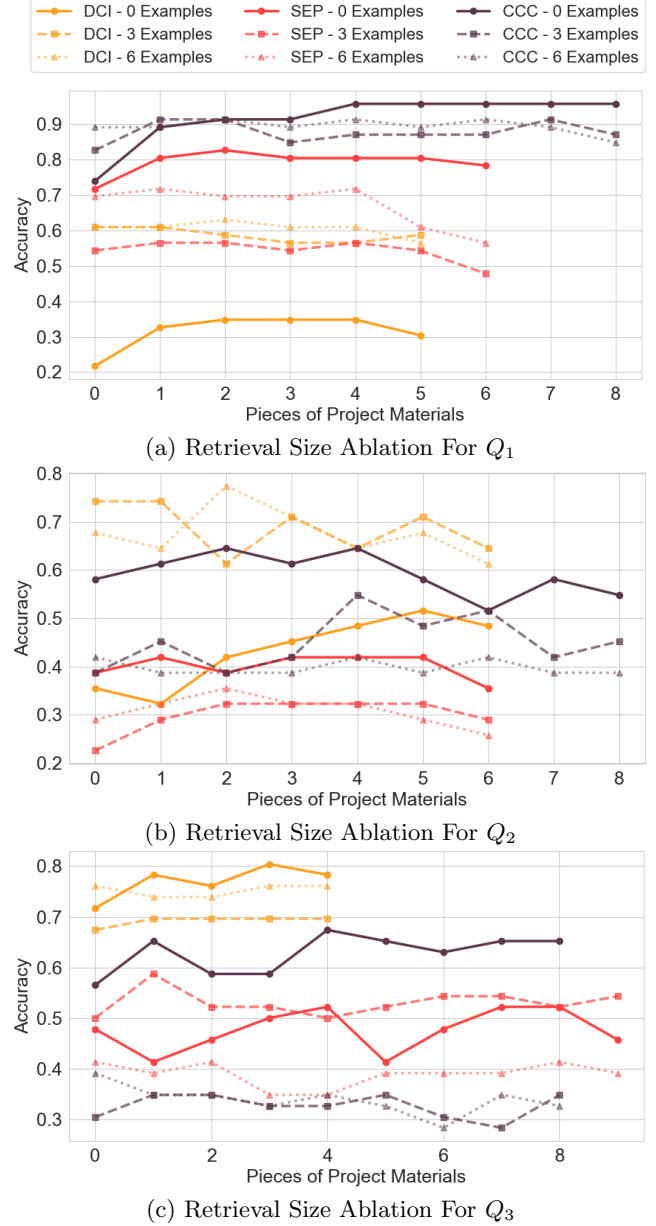

(a) Retrieval Size Ablation For Q_1

(b) Retrieval Size Ablation For Q_2

(c) Retrieval Size Ablation For Q_3

Figure 4: The impact of the retrieval size (k).

vary the retrieval size k from 0 to the maximum available chunks. For Q_1, increasing k in zero-shot settings generally improves zero-shot performance across all dimensions, with CCC showing the most drastic gains. For Q_2, performance typically peaks at moderate k values ($k=2$ or 3) before beginning to decline. Q_3 shows more results, with DCI benefiting from increasing k up to 3 in zero-shot settings, while SEP achieves the best performance at $k=1$ under 3-shot settings and CCC shows optimal results with moderate k values in zero-shot settings. Across all three questions, we observe that the optimal retrieval size k decreases as more examples are provided. In zero-shot scenario, $k=4$ generally leads to the best performance; while $k=1$ typically performs the best in 3-shot scenarios and $k=2$ the best in 6-shot scenarios. This suggests that complementary information can be provided by retrieved knowledge or in-context examples.

5. CONCLUSION

This study investigates into the effectiveness of GradeRAG, a novel framework that integrates retrieval-augmented generation into automatic short answer grading for science education. Our experiments across three science assessment tasks demonstrate that incorporating domain-specific knowledge through RAG significantly improves scoring accuracy, with consistent performance gains across multiple dimensions of scientific understanding. Our experimental results show several advantages of GradeRAG. First, it effectively enhances the alignment between expert graders and automated grading systems by providing access to relevant domain knowledge. Second, our dual-index strategy and the comprehensive retrieval mechanism ensure that the most pertinent information is retrieved for student responses. Lastly, GradeRAG demonstrates compatibility with in-context learning, exhibiting the complementary relationship between the retrieved knowledge and expert examples, which further enhances grading performance.

6. ACKNOWLEDGMENTS

References

[1] A. S. Beatty et al. Developing assessments for the next generation science standards. 2014.

[2] S. Burrows, I. Gurevych, and B. Stein. The eras and trends of automatic short answer grading. International journal of artificial intelligence in education, 2015.

[3] Y. Chu et al. A llm-powered automatic grading framework with human-level guidelines optimization. arXiv preprint arXiv:2410.02165, 2024.

[4] N. R. Council et al. A framework for k-12 science education: Practices, crosscutting concepts, and core ideas. National Academy of Sciences, 2012.

[5] K. Guu et al. Retrieval augmented language model pre-training. In ICML, pages 3929–3938. PMLR, 2020.

[6] G. Harshavardhan and K. Singh. A rubric-centric approach for automated test correction utilizing rag and fine tuning. In ICTACS, 2024.

[7] S. Hassan, A. A. Fahmy, and M. El-Ramly. Automatic short answer scoring based on paragraph embeddings. IJACSA, 2018.

[8] P. He et al. Predicting student science achievement using post-unit assessment performances in a coherent high school chemistry project-based learning system. Journal of Research in Science Teaching, 2023.

[9] P. He et al. A design framework for integrating artificial intelligence to support teachers' timely use of knowledge-in-use assessments. Uses of Artificial Intelligence in STEM Education, 2024.

[10] P. He, X. Zhai, N. Shin, and J. Krajcik. Applying rasch measurement to assess knowledge-in-use in science education. In Advances in Applications of Rasch Measurement in Science Education. Springer, 2023.

[11] C. Leacock and M. Chodorow. C-rater: Automated scoring of short-answer questions. Computers and the Humanities, 2003.

[12] G.-G. Lee et al. Applying large language models and chain-of-thought for automatic scoring. arXiv preprint arXiv:2312.03748, 2024.

[13] P. Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. NeurIPS, 2020.

[14] H. Li et al. Llm-based automated grading with human-in-the-loop. arXiv preprint arXiv:2504.05239, 2025.

[15] Q. Li et al. Adapting large language models for education: Foundational capabilities, potentials, and challenges. arXiv preprint arXiv:2401.08664, 2023.

[16] F. Miladi, V. Psyché, and D. Lemire. Comparative performance of gpt-4, rag-augmented gpt-4, and students in moocs. In International Conference on Breaking Barriers with Generative Intelligence, pages 81–92. Springer, 2024.

[17] M. Mohler, R. Bunescu, and R. Mihalcea. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In ACL 2011, 2011.

[18] N. L. States. Next generation science standards: For states, by states. National Academies Press, 2013.

[19] M. A. Sultan, C. Salazar, and T. Sumner. Fast and easy short answer grading with high accuracy. In NAACL, 2016.

[20] K. Sundar, E. Manohar, K. Vijay, and S. Prakash. Revolutionizing assessment: Ai-powered evaluation with rag and llm technologies. In ICSSAS, pages 43–48. IEEE, 2024.

[21] K. Yang et al. Content knowledge identification with multi-agent large language models (llms). In International Conference on Artificial Intelligence in Education, 2024.