Graph RAG for Automated Short Answer Grading with Feedback: Bridging Pedagogical Needs and Technical Capabilities

Guoliang Xu, James Corter
Columbia University, New York, NY, USA
gx2150@tc.columbia.edu, jec34@tc.columbia.edu

Abstract

Automated short answer grading with feedback (ASAG-F) systems currently face challenges in transparency, pedagogical alignment, and cost-effectiveness that limit their real-world deployment. We introduce GraphRAG, a knowledge graph-based retrieval-augmented generation framework that addresses these limitations by grounding all large language model (LLM)-generated feedback and scores in instructor-curated atomic facts, ensuring traceability and verifiability. Using the Short Answer Feedback (SAF) dataset with 31 topics, we evaluate GraphRAG on unseen-question and unseen-answer splits. Our systematic evaluation demonstrates that GraphRAG achieves grading accuracy comparable to vector-based RAG and generally superior to a fine-tuned LLM baseline model while providing more transparent source attribution. Additional findings include: (1) Instructing the LLM to discretize continuous scores to match pedagogical rubrics, such as the 0.25 increments common in SAF, improves grading accuracy; (2) LLM-generated feedback exhibits length-dependent quality variations when unconstrained; prompt-based length control substantially enhances feedback quality and its stability, achieving optimal balance of instructional richness and conciseness; (3) Performance scaling analysis reveals that basic models like GPT-4o-mini offer cost-effective performance, while premium models like Claude-Opus-4 show diminishing returns. These results demonstrate that GraphRAG offers a robust, explainable, pedagogy-aligned, and cost-effective solution for large-scale educational applications, enabling transparent automated grading with effective pedagogical feedback and practical deployment costs.

Introduction

Short-answer questions assess not only factual recall but also higher-order thinking such as analysis and critical evaluation (Chamberlain et al. 2004). However, manual grading is expertise-intensive, time-consuming, and prone to inconsistency between raters, posing significant scalability challenges. To address these limitations, research on automated assessment has emerged, spanning Automated Essay Scoring (AES), which evaluates holistic aspects (Ramesh and Sanampudi 2022; Yang et al. 2024), and Automated Short Answer Grading (ASAG), our focus, which requires fine-grained semantic and factual accuracy (Burrows, Gurevych and Stein 2015; Madnani and Cahill 2018).

Early ASAG systems relied on rule-based and classical machine learning approaches with engineered features like keyword overlap (Nielsen et al. 2008; Zesch, Wojatzki, and Scholten-Akoun 2015; Galhardi et al. 2024), but they struggled to recognize the full semantic diversity and conceptual equivalence in student language. To address this semantic gap, the adoption of deep learning models (e.g., Convolutional Neural Network, Long Short-Term Memory, and Transformers) marked a significant advance, improving grading accuracy through superior semantic modeling (Taghipour and Ng 2016; Alikaniotis, Yannakoudakis, and Rei 2016; Tay et al. 2018; Liu and Ding 2021; Fowler et al. 2021; Haller 2022; Putnikovic and Jovanovic 2023). This progress, however, came at the cost of transparency and adaptability. The models' black-box nature hindered educator trust, while their reliance on large, domain-specific training datasets limited their scalability (Rudin and Radin 2019).

The advent of Large Language Models (LLMs) marked a paradigm shift. Their powerful zero-shot or few-shot capabilities effectively address the domain adaptation and scalability challenges that hampered previous models (Wei et al. 2022; Huang et al. 2023; Zeng et al. 2023; Zhao et al. 2023). This breakthrough has accelerated AI adoption across educational contexts (Dan et al. 2023; Hackl et al. 2023), including ASAG. Studies in ASAG show LLMs demonstrate proficiency in rubric-based phrase extraction (Yoon 2023) and high self-consistency (Hackl et al. 2023), though cross-domain performance remains somewhat uneven (Schneider et al. 2023; Chang and Ginter 2024). Moreover, LLMs' generative power also enables high-quality, learner-specific feedback generation, advancing the field from ASAG to Automated Short Answer Grading with Feedback (ASAG-F) (Filighera et al. 2022; Li et al. 2023; Schneider et al. 2023). Research shows that feedback is essential for learning gains (Hattie and Timperley 2007; Shute 2008; Li et al. 2023).

However, LLM deployment raises data privacy concerns, because fine-tuning of LLM exposes sensitive student information (Yan et al. 2023). Retrieval-Augmented Generation (RAG) addresses this by retrieving information from secure databases to ground LLM generation without model fine-tuning (Lewis et al. 2020; Khattab and Zaharia 2020). Recent work has applied RAG to ASAG-F (Fateen et al. 2024), demonstrating RAG's viability for ASAG-F. Still, RAG based on vector embeddings links outputs only to large text chunks, not precise facts used, so reasoning remains opaque. Therefore, even with RAG, three practical issues impede deployment of LLM-based ASAG-F in real classrooms:

- Transparency and Verifiability: Current systems, including those enhanced by RAG, remain "black boxes", preventing educators from tracing the rationale behind scoring and feedback, thus hindering trust.
- Alignment with Pedagogical Practices: A significant mismatch persists between model outputs and classroom needs. (1) Models generate continuous scores while classrooms require discrete, rubric-aligned grades (Filighera et al. 2022). (2) Educators need feedback varying from concise hints to detailed explanations without compromising quality. It is unclear if current models preserve instructional quality in generating feedback of different lengths. (3) we hypothesize that constraining feedback length may enhance instructional value by compelling models to focus on critical information, a possibility still unexplored.
- Cost-Effectiveness and Scalability: A focus on computationally expensive LLMs overlooks the practical needs of diverse educational institutions. The performance trade-offs of using smaller, less costly models are unknown.

To address these challenges, we introduce a knowledge graph-based retrieval-augmented generation (GraphRAG) framework for ASAG-F. Unlike traditional RAG, GraphRAG grounds all outputs in an instructor-curated knowledge graph composed of discrete, atomic facts (nodes and edges). By constraining the LLM to retrieve and reason over specific subgraphs, our framework ensures each judgment is directly traceable to the provided facts used, dramatically improving pedagogical transparency and verifiability. This study evaluates GraphRAG 's performance relative to scoring granularity, feedback-length control, and model scale, framing the following five research questions (RQs):

RQ1: Can a GraphRAG system achieve grading accuracy and feedback quality comparable to vector-based RAG or fine-tuned baselines while providing verifiable provenance?

RQ2: How does discretizing continuous AI scores via prompts to align with rubrics affect grading accuracy?

RQ3: Does feedback quality hold steady as the model generates shorter versus longer explanations?

RQ4: Can prompt-based length constraints be used to improve performance by focusing on relevant content?

RQ5: How robust and scalable is the performance of our GraphRAG framework as the capability (and cost) of the backbone LLM increases?

Contributions: (1) We design the first GraphRAG framework for ASAG-F that uses knowledge graphs to ensure all LLM-generated grades and feedback are fully traceable to instructor-approved sources. (2) We conduct the first systematic analysis of the interplay between scoring granularity and accuracy, showing that discretizing continuous LLM scores to match rubric-based intervals can significantly improve grading agreement. (3) We identify length-dependent quality variations in LLM-generated feedback and provide the first empirical evidence that prompt-based length constraints in the GraphRAG framework can mitigate this instability, enhancing both instructional value and grading consistency. (4) We demonstrate the robustness of the GraphRAG framework across different LLMs, enabling cost-effective deployment in real-world educational settings.

Related Work

LLM-based ASAG: Promise and Barriers

Recently, LLMs have revolutionized ASAG through two key capabilities. First, they demonstrate strong scoring performance with high self-consistency (Hackl et al. 2023) and rubric-based proficiency (Yoon 2023), though alignment with human graders is only moderate across subjects (Schneider et al. 2023; Chang and Ginter 2024). Second, they enable high-quality feedback generation, advancing the field from template-based ASAG systems (Riordan et al. 2017; Sung et al. 2019) to dynamic ASAG-F systems that align better with pedagogical needs for student feedback (Hattie and Timperley 2007; Li et al. 2023; Naismith et al. 2023; Yancey et al. 2023). However, these advances face two fundamental barriers limiting practical classroom deployment. First, opacity: LLM-based ASAG-F remains a "black box", limiting trust even with post-hoc explanations (Tornqvist et al. 2023). Second, significant data privacy concerns arise, because fine-tuning or pretraining with student data can expose sensitive information (Yan et al. 2023).

RAG: Addressing Privacy Concerns

To address privacy and grounding issues, Retrieval-Augmented Generation (RAG) has emerged as a state-of-the-art method for LLMs (Lewis et al. 2020; Wang et al. 2025). In ASAG-F, privacy data are embedded into secure vector databases, which are used to retrieve semantically similar exemplars to guide LLM generation without direct fine-tuning (Gao et al. 2024; Fateen et al. 2024). Despite effectively mitigating privacy concerns, RAG (vector-based retrieval) often lacks verifiable grounding, as retrieved results are stored as dense, opaque embeddings (Hu and Lu 2024).

GraphRAG: RAG with Knowledge Graphs

To address this limitation, recent work explores replacing v-

ector retrieval with Knowledge Graphs (KGs), which encode information as discrete, symbolic (subject, predicate, object) triples (Ji et al. 2022). This symbolic structure enables precise attribution to specific, human-readable facts, thereby enhancing verifiability (Abu-Salih 2021; Pan et al. 2024; Procko 2024; Tian et al. 2025). Building on this idea, GraphRAG integrates KGs into RAG by retrieving relevant subgraphs to constrain LLM generation, reducing hallucinations and improving factual accuracy while providing interpretable reasoning paths (Chen et al. 2024; Zhu et al. 2024; Procko 2024). This emphasis on explicit structure aligns with broader trends in improving LLM reasoning through graph-based methods like Tree of Thoughts and Graph of Thoughts (Yao et al. 2024; Besta et al. 2024). GraphRAG has demonstrated substantial gains across knowledge-intensive applications (Wu et al. 2023; Feng, Zhang, and Fei 2023; Ding et al. 2024) despite remaining challenges with domain-specific biases (Caufield et al. 2024). The deployment of GraphRAG is now more accessible due to advances in automated KG construction via LLM-based extraction (Safavi and Koutra 2021; Zhang et al. 2024) and enterprise systems like Microsoft's GraphRAG (Edge et al. 2024). Despite proven value in other domains, its application to educational assessment and ASAG-F is currently nascent.

Methods

Our GraphRAG framework for transparent ASAG-F operates in three phases: (1) data preprocessing (2) automated construction of a pedagogical Knowledge Graph (KG) from reference materials, and (3) an LLM-retrieval-generation pipeline that scores student answers and provides traceable feedback using the KG. All materials are archived in Zenodo.

Data

We use the English Short Answer Feedback (SAF) dataset (Filighera et al. 2022), covering 31 college-level topics in communication networks. SAF suits ASAG-F because it pairs numeric scores with content-focused human feedback, and it has been adopted in recent RAG-based ASAG work (e.g., Fateen et al. 2024), making it a practical benchmark. Each instance contains: a question; a reference answer; a student answer; numeric score (0.0-1.0); and detailed human-written feedback (our gold standard). The training set (1,700 instances) is used for KG construction. For evaluation, we use two cleaned (removing several cases with score >1), held-out test sets: unseen answers (313 instances) to test robustness to new phrasings, and unseen questions (406 instances) to test generalization to new topics.

Data Preprocessing

The raw dataset contains a flat list of individual student responses without direct linkage between answers related to t-

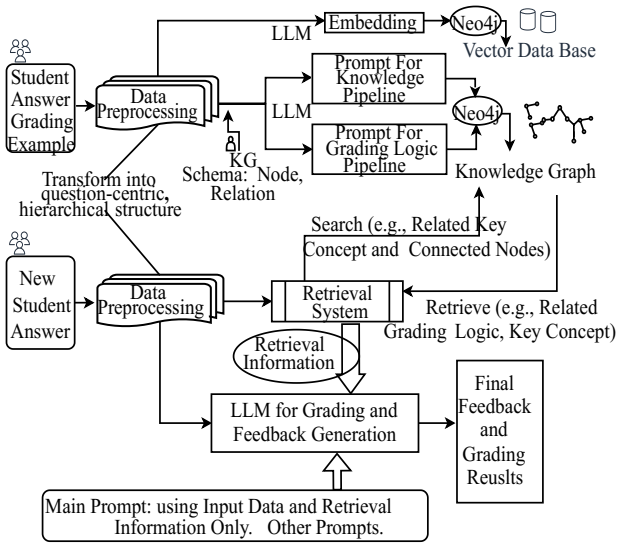

Figure 1: An Overview of the GraphRAG Framework for ASAG-F.

he same question. To facilitate generalization, we transformed the flat data into a question-centric, hierarchical structure (cf. Chen et al. 2023), grouping the entire dataset by unique question text (Figure 1). The result is a collection of data "bundles," where each bundle contains all information pertinent to a single question: question text, reference answer(s), and all student responses with their scores and feedback. This question-centric bundling enables the model to analyze the full spectrum of answers from correct to partially correct to common mistakes, thereby constructing a far richer and more accurate KG, and to provide comprehensive context during inference for new answers.

Knowledge Graph Construction

The next step is to distill the unstructured text data into a structured, queryable graph that captures both domain knowledge and assessment logic. We work entirely in Neo4j and proceed in two phases: schema design and population.

KG Schema

To ensure that the KG is both structured and meaningful, we define a domain-specific schema for it, consisting of predefined node labels and relation types. Node labels represent key educational entities, spanning essential disciplinary concepts, structural components of answers, common student mistakes, and rubric-based evaluation criteria. Relation types capture the instructional, logical, and semantic dependencies among these entities, such as prerequisites, evaluation links, compositional and evidential connections, diagnostic cues, and mappings from issues to remedial suggestions. During inference, criterion-level evidence is aggregated along admissible, typed paths to generate scores,

and feedback is derived by traversing diagnostic links toward corrective suggestions.

KG Population Pipeline

The process of populating the KG from the prepared text bundles is automated through a comprehensive pipeline. For each data bundle, the pipeline executes four steps:

(1) Text Chunking: The source texts (e.g., a reference answer in training datasets) are segmented into semantically coherent chunks for further manageable LLM processing.

(2) LLM-based Information Extraction: Each chunk is processed with GPT-4o-mini, guided by the KG schema in two prompt-guided processing pipelines:

- Knowledge Pipeline: uses a knowledge prompt to process reference answers to extract foundational domain knowledge and relationships, building the course ontology representing expected student knowledge into the KG using symbolic triples.
- Grading Logic Pipeline: uses a grading prompt to analyze student answers, score and feedback to capture the assessment reasoning, learning why answers lose points and how effective feedback is constructed.

Both pipelines extract entities and relationships as structured JSON outputs following the predefined KG schema, which ensures that our KG is not merely a repository of facts, but a rich model embodying both the core subject matter and nuanced evaluation logic. Use of GPT models for automated KG construction has been validated in prior work (Zhu et al. 2024). Following best practices in prompt engineering (Liu et al. 2023), detailed prompts are in the archived Appendix.

(3) Graph Ingestion: The extracted structured outputs (JSON) are parsed into a Neo4j graph via Cypher, ensuring deduplication and semantic connectivity across nodes.

(4) Embedding and Indexing: Each chunk node's text is embedded into a vector using OpenAI's text-embedding-3-small model for its balance of speed and accuracy. These embeddings are stored as node properties and indexed in Neo4j to enable efficient similarity search during inference.

Retrieval Frameworks and Generation via Prompt Engineering

Retrieval Frameworks

At inference, new answers are reorganized into question-centric bundles (as in training) to ensure consistent processing and formulate retrieval queries (input for retrieval). We implement and compare two retrieval frameworks, both paired with GPT-4o (Hurst et al. 2024), for final generation:

- Vector-RAG (RAG Baseline): Dense retrieval computing cosine similarity between inference bundles and embedded chunks, returning top-matching passages as context.
- Graph-RAG (Ours): Two-step Vector-Cypher process: (1) vector search identifies top "seed" nodes in the KG; (2) Cypher traversal retrieves 1–2-hop neighboring nodes. Context combines text from the top 3 seeds and top 2 neighbors, yielding richer, structured retrieval.

The main distinction lies in evidentiary granularity: Vector-RAG returns unstructured passages with coarse traceability, whereas Graph-RAG returns a subgraph of explicitly linked facts, enabling fine-grained, transparent attribution.

Generation via Prompt Engineering

The context retrieved by either framework feeds into the generation stage, guided by a structured prompt that functions as a reasoning scaffold for the LLM (here, we use GPT-4o). The prompt's architecture enforces several key principles: assigns the LLM the role of an "assessment expert," requires a structured JSON output for scores and feedback, and applies an evidence-based constraint requiring all LLMs' judgments to rely solely on the provided context. This mitigates hallucination and ensures traceability.

Experimental Setup

We evaluate our GraphRAG framework for ASAG-F on two tasks: (1) Grading Accuracy and (2) Generated Feedback Quality, directly aligned with our research questions.

Models and Baselines

We compared three distinct systems on the SAF dataset (Filighera et al. 2022) to assess benefits of our approach:

- GraphRAG (Ours): a pedagogical knowledge graph built via a dual-pipeline strategy and retrieved using a graph-based Retriever: Vector-Cypher (detailed in Method).
- Vector-RAG (RAG Baseline): a strong baseline, Vector-RAG method, proposed by Fateen et al. (2024). It utilizes an identical LLM (gpt-4o) and prompts as our system but relies on standard dense vector retrieval.
- Fine-tuned Transformer (Fine-tuned LLM Baseline): T5-based model fine-tuned on the SAF dataset, as provided by the dataset's creators (Filighera et al. 2022), representing a powerful, non-RAG paradigm.

Evaluation Scenarios

All experiments are conducted on the Unseen Answers and Unseen Questions splits of the SAF dataset to test different generalization capabilities. To address the research questions, we conduct several targeted studies:

- Overall Performance: Compare GraphRAG, VectorRAG, and fine-tuned baseline on grading and feedback quality.
- Scoring Granularity: Prompt LLMs to produce scores at discrete intervals (increments of 0.1, 0.125, 0.2 or 0.25) to simulate real-world rubric requirements and assess accuracy impact.
- Feedback Length and Quality Relationship: Group naturally generated feedback into length bins (0-9, 10-19, 20-39 words) and compute the mean BERTScore for each bin to explore the relationship between length and quality.

- Length Control: Apply prompt-based constraints (none, ≤30 words, ≤50 words) on feedback and evaluate quality across all pedagogical metrics.
- Model Scaling: Compare different LLM generators' performance (GPT-4o, Claude-Opus-4, GPT-4o-mini).

Evaluation Metrics

We evaluate two tasks using distinct metrics. Grading (Task 1): Pearson Correlation and RMSE for score alignment; Exact Match rates and Tolerance Match rates (±0.2, ±0.25) for practical agreement. Generated Feedback Quality (Task 2): BERTScore for semantic similarity to human feedback and four key pedagogical metrics assessing instructional value: (1) Diagnostic Accuracy, proportion of knowledge gaps (concepts present in the reference answer but missing from the student answer) that are correctly identified in generated feedback, (2) Topical Alignment, the embedding cosine similarity between generated feedback and the reference answer, measuring how on-topic the feedback is, (3) Actionable Verbs, count of instructional verbs (e.g., "explain," "suggest," "analyze") drawn from the pedagogical lexicon to evaluate instructional utility, (4) Keyword Overlap, percentage of reference answer key terms present in feedback, to assess coverage of essential terminology.

Implementation Details

The described system was built in Python with Neo4j for the graph database. We used GPT-4o for quality-sensitive generation, cost-efficient GPT-4o-mini for KG construction, and text-embedding-3-small for vectorization. Temperature = 0.0 ensures determinism.

Model	Corr	RMSE	EMR	TM2	TM25
Unseen Answers					
Baseline	0.22	0.39	**0.65**	0.71	**0.81**
Vector-RAG	0.73	**0.22**	0.15	**0.73**	0.75
Graph-RAG	**0.74**	**0.22**	0.15	**0.73**	0.74
Unseen Questions					
Baseline	0.13	0.42	**0.44**	0.47	0.66
Vector-RAG	**0.79**	**0.24**	0.10	**0.68**	**0.73**
Graph-RAG	**0.79**	0.25	0.08	0.63	0.68

Table 1: Autograding Performance for Vector RAG, GraphRAG, and Pre-trained (fine-tuned) Baseline Models, on Unseen Data. EMR = Exact Match Rate, TM2 = Tolerance Match Rate (±0.2), TM25=Tolerance Match Rate (±0.25).

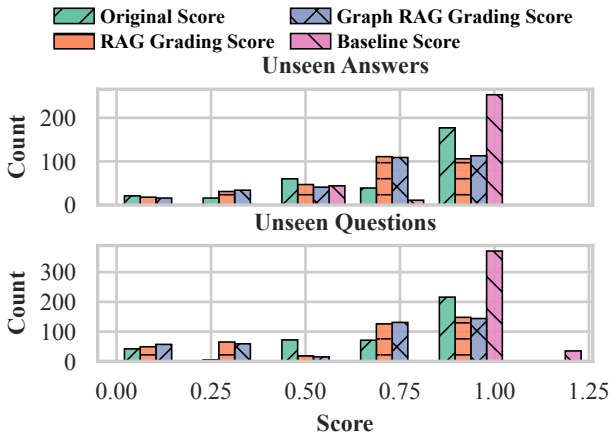

Figure 2: Distributions of Original Scores Compared to Scores from LLM Baseline, Vector-RAG and GraphRAG Models on Unseen Answers and Unseen Questions.

Results and Discussion

Overall Performance Comparison

To answer RQ1, we compare GraphRAG with Vector-RAG and fine-tuned baseline models on grading accuracy and feedback quality. For grading accuracy, across both splits, GraphRAG matches Vector-RAG on correlation and RMSE, with both vastly outperforming the fine-tuned baseline (Table 1). When scores are evaluated categorically, the baseline model surprisingly has the highest exact match rate (EMR), despite its very low score correlation. As Figure 2 shows, the baseline model is predominantly assigning scores of 1 for the unseen answers and almost entirely scores ≥1 for unseen questions. Essentially, the baseline model is implementing a "maximizing" strategy that predicts the most common score. Such scoring is not useful pedagogically, despite the high EMR. In contrast, both RAG models generate realistic score distributions more closely aligned with human assessments. Regarding feedback quality (Table 2), the baseline model achieves the highest BERTScore for unseen answers but is comparable to other models for unseen questions. Feedback text from RAG models tends to be longer, and scores higher on the feedback pedagogical quality measures like diagnostic accuracy (GraphRAG's 0.127 versus baseline's 0.039 on Unseen Answers). That is, RAG models generate longer, instructionally richer feedback with more actionable verbs and keyword coverage, confirming the pedagogical usefulness. Crucially, GraphRAG matches Vector-RAG performance while uniquely providing verifiable source attribution through its KG structure**.**

Effect of Grading Granularity and Rubric Alignment on Grading Performance

For RQ2, we test how discretizing continuous AI scores via prompt constraints impacts grading performance, when aligned with common pedagogical rubrics. Figure 2 shows human grades appear to be based on subdividing the full 1-point scale into 2, 4 or occasionally 8 equal parts, clustering around increments of 0.25 and 0.125 (e.g., 0.5, 0.75, 0.875). When we constrain the AI's output to these same increments,

Dataset	Model	BERTScore	Average Length	Topical Alignment	Keyword Overlap	Actionable Verbs (Avg)	Diagnostic Accuracy
Unseen Answers	Human FB	1.000	19.3 words	0.425	0.038	0.173	0.041
	Baseline	**0.944**	13.6 words	0.398	0.031	0.077	0.039
	Vector-RAG	0.859	71.3 words	**0.682**	**0.151**	**0.825**	**0.120**
	Graph-RAG	0.859	72.4 words	**0.683**	**0.153**	**0.788**	**0.127**
Unseen Questions	Human FB	1.000	29.5 words	0.440	0.027	0.219	0.031
	Baseline	**0.898**	10.2 words	0.246	0.013	0.054	0.008
	Vector-RAG	0.870	75.4 words	**0.682**	**0.084**	**0.585**	**0.086**
	Graph-RAG	0.870	75.2 words	**0.686**	**0.084**	**0.665**	**0.085**

Table 2: Feedback Quality Measures for Basic Vector-RAG, Basic GraphRAG, and the Fine-tuned LLM Baseline, on Unseen Questions and Unseen Answer.

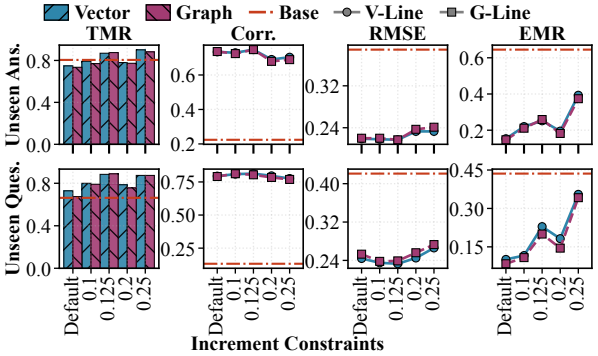

Figure 3: Effects of Varying Discretization Constraints on Grading Performance for Vector-RAG and GraphRAG.

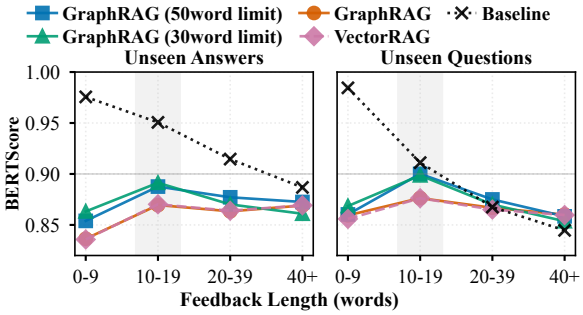

Figure 4: Feedback Quality (as Measured by BERTScore) Across Different Feedback Length Bins.

it directly boosts the GraphRAG system's performance on agreement-based metrics (see Figure 3), with the 0.25 (or 0.125) increment yielding a 75% improvement in the Exact Match Rate (EMR) and the highest Tolerance Match Rate (TMR at ±0.25). RMSE shows a shallow U-shape, minimizing at 0.1-0.125 but increasing at coarser granularities (0.2-0.25). This suggests that while coarser increments may improve rubric-level agreement, they can introduce larger errors in individual predictions, reflecting the trade-off between fidelity and noise reduction. Correlation is stable (~0.8) across all granularities, preserving rank order.

These findings empirically demonstrate that discretizing AI Scores to match rubric intervals can increase the agreement with human scores without sacrificing correlation and ranking quality. To our knowledge, this "alignment effect" in automated assessments has not been identified elsewhere.

Feedback Quality and Length Trade-offs

To addresses RQ3, we examine feedback quality across different generation lengths. Figure 4 reveals significant length-dependent quality variations in uncontrolled generation. The baseline exhibits substantial degradation (9-12% BERTScore drop) as length increases, while GraphRAG and Vector-RAG show a pronounced peak at 10-19 words (0.890), aligning with typical human feedback length.

Quality diminishes at both extremes (<10 words: 0.836; >40 words: 0.860), creating pedagogical challenges: students may receive inconsistent feedback quality based solely on generation length. These findings motivate exploration of explicit length control (RQ4).

Effect of Prompt-Based Length Control on Grading Performance and Feedback Quality

Next, we test RQ4: whether prompt-based length constraints boost grading accuracy, feedback quality and instructional value by focusing models on relevant content. Length constraints prove remarkably effective, transforming both grading accuracy and feedback quality (Tables 3-4; Figure 4). With the 50-word constraint, correlation improves from 0.74 (unconstrained) to 0.76, 0.25-tolerance match rate from 0.74 to 0.85, and BERTScore from 0.859 to 0.875. The 30-word variant maintains comparable performance, confirming that constraints enhance rather than compromise quality.

Most importantly, constrained GraphRAG surpasses human feedback on pedagogical quality metrics (Table 3). The 50-word configuration doubles human diagnostic accuracy (0.083 vs. 0.041) and provides more actionable guidance (0.240 vs. 0.173 actionable verbs), with superior keyword coverage and topical alignment. This suggests length constraints help models focus on core instructional content that humans sometimes miss.

Scenario	Model	BERT Score	Average Length	Topical Alignment	Keyword Overlap	Actionable Verbs (avg)	Diagnostic Accuracy
unseen answers	Human Feedback	1.000	19.3 words	0.425	0.038	0.173	0.041
	Baseline	0.944	13.6 words	0.398	0.031	0.077	0.039
	Graph-RAG	0.859	72.4 words	0.683	0.153	0.788	0.127
	Graph-RAG-30words Limits	0.875	14.7 words	0.467	0.057	0.137	0.050
	Graph-RAG-50words Limits	0.875	28.1 words	0.595	0.091	0.240	0.083
unseen questions	Human Feedback	1.000	29.5 words	0.440	0.027	0.219	0.031
	Baseline	0.898	10.2 words	0.246	0.013	0.054	0.008
	Graph RAG	0.870	75.2 words	0.686	0.084	0.665	0.085
	Graph-RAG-30words Limits	0.883	15.0 words	0.461	0.035	0.189	0.033
	Graph-RAG-50words Limits	0.880	29.7 words	0.574	0.043	0.226	0.047

Table 3: Feedback Quality Metrics Across Different Length Control Conditions on Unseen Data.

Model	Corr	RMSE	EMR	TM2	TM25
Unseen Answers					
Baseline	0.22	0.39	**0.65**	0.71	0.81
GraphRAG	0.74	0.22	0.15	0.72	0.74
GraphRAG-30w	**0.76**	0.21	0.28	**0.81**	**0.85**
GraphRAG-50w	**0.76**	**0.20**	0.27	**0.81**	**0.85**
Unseen Questions					
Baseline	0.13	0.42	**0.44**	0.47	0.66
Graph-RAG	0.79	0.25	0.08	0.63	0.68
Graph-RAG-30w	**0.80**	**0.24**	0.21	**0.75**	**0.82**
Graph-RAG-50w	**0.80**	0.25	0.19	0.74	**0.82**

Table 4: Effect of Prompt-Based Length Control on Grading Performance, for Unseen Data. EMR = Exact Match Rate, TM2/TM25 = Tolerance Match Rate (±0.2/±0.25), 30w/50w Means 30/50-Word limitations on feedback.

Although the fine-tuned LLM baseline achieves highest BERTScore (0.944) through minimal feedback (13.6 words), its instructional value remains poor (diagnostic accuracy = 0.039; actionable verbs = 0.077). In contrast, GraphRAG with a 50-word limit (BERTScore: 0.875) seems to achieve the best overall trade-off, matching human-like response length while providing more focused, actionable, and accurate feedback than either baseline model or human responses. These improvements generalize across unseen questions, confirming that length control is a simple yet powerful enhancement for educational AI systems.

Performance-Cost Tradeoff

For RQ5, Table 5 reveals that quality plateaus with the cheapest model. BERTScore barely changes across GPT-4o-mini to Claude-Opus-4 (~0.84-0.85), yet cost-efficiency plummets 60-fold (TM2/Cost in USD/Mtok: 2.47 to 0.04). GPT-4o optimizes performance-cost balance while GPT-4o-mini dominates efficiency metrics. Claude-Opus-4's marginal gains cannot justify its premium pricing. This plateau effect means that schools with budget constraints can deploy GraphRAG without sacrificing quality, democratizing access to high-quality automated assessment.

Model	TM2	BERT Score	TM2/ Cost	BERT/ Cost
4o-mini-Vector	0.30	**0.85**	1.98	**5.66**
4o-mini-Graph	0.37	**0.85**	**2.47**	**5.66**
4o-Vector	**0.59**	**0.85**	0.24	0.34
4o-Graph	0.56	**0.85**	0.22	0.34
Opus-4-Vector	0.52	0.84	0.04	0.06
Opus-4-Graph	0.52	0.84	0.04	0.06

Table 5: RAG Performance-Cost Analysis for Different Backbone LLMs. TM2: Tolerance Match Rate (±0.2).

Conclusions

We describe the design and implementation of the first GraphRAG framework for ASAG-F that uses knowledge graphs to ensure all AI-generated grading and feedback are verifiably linked to instructor-approved knowledge, providing traceability. Compared with a fine-tuned LLM baseline, GraphRAG reduces grade inflation, closely matching human score distributions, and produces feedback that scores higher on key quality indicators. Our findings can directly benefit educational practice: educators can achieve better AI-human grading consistency while preserving relative student rankings by constraining AI to use rubric intervals; uncontrolled feedback generation produces unstable quality across different lengths; however, implementing feedback length constraints through AI prompts ensures both stability and higher instructional quality, offering clear implementation guidance; the GraphRAG system is effective with minimal-cost LLMs, making deployment feasible for budget-constrained institutions. These contributions can help make high-quality automated grading and feedback accessible in diverse educational settings, ultimately enhancing learning outcomes through consistent, explainable, and timely feedback. Future work can examine domain generalizability and whether length constraints emulate cognitive load regulation.

References

Abu-Salih, B. 2021. Domain-specific Knowledge Graphs: A survey. arXiv:2011.00235.

Alikaniotis, D.; Yannakoudakis, H.; and Rei, M. 2016. Automatic text scoring using neural networks. *arXiv preprint arXiv*:1606.04289.

Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawski, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nyczyk, P.; et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 17682–17690.

Burrows, S.; Gurevych, I.; and Stein, B. 2015. The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education*, 25(1): 60–117.

Caufield, J. H.; Hegde, H.; Emonet, V.; Harris, N. L.; Joachimiak, M. P.; Matentzoglu, N.; Kim, H.; Moxon, S.; Reese, J. T.; Haendel, M. A.; et al. 2024. Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning. *Bioinformatics*, 40(3): btae104.

Chamberlain, C.; Button, A.; Dison, L.; Granville, S.; and Delmont, E. 2004. The role of short answer questions in developing higher order thinking. *Per Linguam: a Journal of Language Learning Tydskrif vir Taalaanleer*, 20(2): 28–45.

Chang, L.-H.; and Ginter, F. 2024. Automatic short answer grading for finnish with chatgpt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 23173–23181.

Chen, J.; Liu, Z.; Huang, S.; Liu, Q.; and Luo, W. 2023. Improving interpretability of deep sequential knowledge tracing models with question-centric cognitive representations. In *Proceedings of the AAAI conference on artificial intelligence*, 14196–14204.

Chen, J., Xiao, S., Zhang, P.; et al. 2024. BGE M3- Embedding: Multi-Lingual, Multi-Functionality, Multi- Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv:2402.03216.

Dan, Y.; Lei, Z.; Gu, Y.; Li, Y.; Yin, J.; Lin, J.; Ye, L.; Tie, Z.; Zhou, Y.; Wang, Y.; Zhou, A.; Zhou, Z.; Chen, Q.; Zhou, J.; He, L.; and Qiu, X. 2023. EduChat: A Large-Scale Language Model-based Chatbot System for Intelligent Education. arXiv:2308.02773.

Ding, Y.; Fan, W.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.-S.; and Li, Q. 2024. A survey on rag meets llms: Towards retrieval-augmented large language models. *arXiv preprint arXiv*:2405.06211.

Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; and Larson, J. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv*:2404.16130.

Fateen, M.; Wang, B.; and Mine, T. 2024. Beyond scores: A modular rag-based system for automatic short answer scoring with feedback. *IEEE Access*.

Feng, C.; Zhang, X.; and Fei, Z. 2023. Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs. *arXiv preprint arXiv*:2309.03118.

Filighera, A.; Parihar, S.; Steuer, T.; Meuser, T.; and Ochs, S. 2022. Your answer is incorrect... would you like to know why? introducing a bilingual short answer feedback dataset. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 8577–8591.

Fowler, M.; Chen, B.; Azad, S.; West, M.; and Zilles, C. 2021. Autograding "Explain in Plain English" questions using NLP. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, SIGCSE '21, 1163–1169. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380621.

Galhardi, L.; Herculano, M.F.; Rodrigues, L.; Miranda, P.; Oliveira, H.; Cordeiro, T.; Bittencourt, I.I.; Isotani, S.; and Mello, R.F. 2024. Contextual Features for Automatic Essay Scoring in Portuguese. In *International Conference on Artificial Intelligence in Education*, 270-282. Springer.

Gao, Y., Xiong, Y., Gao, X.; et al. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997.

Hackl, V.; Müller, A. E.; Granitzer, M.; and Sailer, M. 2023. Is GPT-4 a reliable rater? Evaluating Consistency in GPT-4 Text Ratings. arXiv:2308.02575.

Haller, S.; Aldea, A.; Seifert, C.; and Strisciuglio, N. 2022. Survey on Automated Short Answer Grading with Deep Learning: from Word Embeddings to Transformers. arXiv:2204.03503.

Hattie, J.; and Timperley, H. 2007. The power of feedback. *Review of educational research*, 77(1): 81–112.

Huang, S.; Dong, L.; Wang, W.; Hao, Y.; Singhal, S.; Ma, S.; Lv, T.; Cui, L.; Mohammed, O.K.; Patra, B.; and Liu, Q. 2023. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36: 72096–72109.

Hu, Y.; and Lu, Y. 2024. Rag and rau: A survey on retrieval-augmented language model in natural language processing. *arXiv preprint arXiv*:2404.19543.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv*:2410.21276.

Ji, S., Pan, S., Cambria, E.; et al. 2022. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2): 494–514.

Khattab, O.; and Zaharia, M. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 39–48.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.T.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 9459–9474. Curran Associates, Inc.

Li, Z.; Zhang, C.; Jin, Y.; Cang, X.; Puntambekar, S.; and Passonneau, R. J. 2023. Learning When to Defer to Humans for Short Answer Grading. In *International Conference on Artificial Intelligence in Education*, 414–425. Springer.

Liu, C.; and Ding, G. 2021. MFDNN: Mixed Features Deep Neural Network Model for Prompt-independent Automated Essay Scoring. In *Proceedings of the 2021 4th International Conference on Algorithms, Computing and Artificial Intelligence*, 1–7.

Liu, Y.; Deng, G.; Xu, Z.; Li, Y.; Zheng, Y.; Zhang, Y.; Zhao, L.; Zhang, T.; Wang, K.; and Liu, Y. 2023. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv*:2305.13860.

Madnani, N.; and Cahill, A. 2018. Automated Scoring: Beyond Natural Language Processing. In Bender, E. M.; Derczynski, L.; and Isabelle, P., eds., *Proceedings of the 27th International Conference on Computational Linguistics*, 1099–1109. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Naismith, B.; Mulcaire, P.; and Burstein, J. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Workshop on Innovative Use of NLP for Building Educational Applications*.

Nielsen, R. D.; Ward, W.; Martin, J.; and Palmer, M. 2008. Annotating Students' Understanding of Science Concepts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA).

Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; and Wu, X. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Procko, T. 2024. Graph Retrieval-Augmented Generation for Large Language Models: A Survey. *Available at SSRN*.

Putnikovic, M.; and Jovanovic, J. 2023. Embeddings for Automatic Short Answer Grading: A Scoping Review. *IEEE Transactions on Learning Technologies*, 16(2): 219–231.

Ramesh, D.; and Sanampudi, S.K. 2022. An automated essay scoring system: a systematic literature review. *Artificial Intelligence Review*, 55(3): 2495–2527.

Riordan, B.; Horbach, A.; Cahill, A.; Zesch, T.; and Lee, C. 2017. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th workshop on innovative use of NLP for building educational applications*, 159–168.

Rudin, C.; and Radin, J. 2019. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, 1(2): 1–9.

Safavi, T.; and Koutra, D. 2021. Relational world knowledge representation in contextual language models: A review. *arXiv preprint arXiv*:2104.05837.

Schneider, J.; Schenk, B.; Niklaus, C.; and Vlachos, M. 2023. Towards LLM-based Autograding for Short Textual Answers. arXiv:2309.11508.

Shute, V.J. 2008. Focus on formative feedback. *Review of educational research*, 78(1):153–189.

Sung, C.; Dhamecha, T. I.; and Mukhi, N. 2019. Improving short answer grading using transformer-based pre-training. In *International Conference on Artificial Intelligence in Education*, 469–481. Springer.

Taghipour, K.; and Ng, H. T. 2016. A Neural Approach to Automated Essay Scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1882–1891. Austin, Texas: Association for Computational Linguistics.

Tay, Y.; Phan, M. C.; Tuan, L. A.; and Hui, S. C. 2018. SKIPFLOW: Incorporating Neural Coherence Features for End-to-End Automatic Text Scoring. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Tian, S.; Xing, S.; Li, X.; Luo, Y.; Yuan, C.; Chen, W.; Jiang, H.; and Wang, X. 2025. A systematic exploration of knowledge graph alignment with large language models in retrieval augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 25291–25299.

Tornqvist, M.; Mahamud, M.; Guzman, E. M.; and Farazouli, A. 2023. ExASAG: Explainable framework for automatic short answer grading. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, 361–371.

Wang, Y.; Zhang, H.; Pang, L.; Guo, B.; Zheng, H.; and Zheng, Z. 2025. MaFeRw: Query rewriting with multi-aspect feedbacks for retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 25434–25442.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.

Wu, Y., Hu, N., Bi, S.; et al. 2023. Retrieve-Rewrite-Answer: A KG-to-Text Enhanced LLMs Framework for Knowledge Graph Question Answering. arXiv:2309.11206.

Yan, L.; Sha, L.; Zhao, L.; Li, Y.; Martinez-Maldonado, R.; Chen, G.; Li, X.; Jin, Y.; and Gašević, D. 2023. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*.

Yancey, K.; LaFlair, G. T.; Verardi, A.; and Burstein, J. 2023. Rating Short L2 Essays on the CEFR Scale with GPT-4. In *Workshop on Innovative Use of NLP for Building Educational Applications*.

Yang, K.; Raković, M.; Li, Y.; Guan, Q.; Gašević, D.; and Chen, G. 2024. Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability. In *Proceedings of the AAAI conference on artificial intelligence*, 22466–22474.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Yoon, S.-Y. 2023. Short Answer Grading Using One- shot Prompting and Text Similarity Scoring Model. arXiv:2305.18638.

Zeng, Z.; Li, L.; Guan, Q.; Gašević, D.; and Chen, G. 2023. Generalizable Automatic Short Answer Scoring via Proto- typical Neural Network. In *International Conference on Artificial Intelligence in Education*, 438–449. Springer.

Zesch, T.; Wojatzki, M.; and Scholten-Akoun, D. 2015. Task-independent features for automated essay grading. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, 224–232.

Zhang, N.; Yao, Y.; Tian, B.; Wang, P.; Deng, S.; Wang, M.; Xi, Z.; Mao, S.; Zhang, J.; Ni, Y.; et al. 2024. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv*:2401.01286.

Zhao, W. X., Zhou, K., Li, J.; et al. 2023. A Survey of Large Language Models. arXiv:2303.18223.

Zhu, Y.; Wang, X.; Chen, J.; Qiao, S.; Ou, Y.; Yao, Y.; Deng, S.; Chen, H.; and Zhang, N. 2024. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web*, 27(5): 58.