Assessing the Quality and Consistency of Automated Knowledge Component Generation using Instructor-generated Questions and LLMs

Jordan Esiason[1], Priyanka Khare[1], Wookhee Min[1], Seung Lee[1], Gamze Ozogul[2], Xiaoying Zheng[2], Yeil Jeong[2], James Lester [1]

[1]North Carolina State University, [2]Indiana University – Bloomington

Correspondence: jesiaso@ncsu.edu

Abstract

Lecture-style instruction is one of the most prevalent forms of learning in postsecondary education in the United States. Despite the factors that make lectures a convenient format, they tend to present few opportunities for meaningful engagement between students and the course materials being presented due to factors such as the overhead associated with interacting with large numbers of students. By utilizing large language models, we have created a pipeline built upon the EXPLAINIT classroom response system for processing student self-explanations produced during lectures using automatically generated knowledge components. This pipeline can facilitate deeper engagement with course materials, offer traceability in assessment results, and allows instructors to respond to student errors or misconceptions in real-time during lecture. While previous work using a proprietary large language model has examined the basic functionality of this pipeline, this work more closely examines the consistency and quality of this pipeline using both a large closed-weight model and a smaller open-weight model, with or without retrieval-augmented generation (RAG). The use of open-source models could allow institutions deploying EXPLAINIT to maintain control of their student data without substantially sacrificing performance. We find that while there are small statistically significant differences in performance between the RAG conditions of each LLM, they are nearly comparable at this task. Additionally, the LLM-generated knowledge components are of higher quality when relevant course material is provided for RAG, although consistency is not improved. These results indicate that both large closed-weight and smaller open-weight models show promise in this task, but fine-tuning may be necessary to improve performance further.

1 Introduction

Lecture-style instruction is one of the most popular teaching formats in the United States due to factors such as course size and the availability of appropriate classrooms (Stains et al., 2018). However, this style of lecture suffers from drawbacks such as a lack of interactivity that can lead to poorer learning outcomes compared to students in more interactive (or dialogic) learning environments (Grissom et al., 2017). Classroom response systems (CRSs) may represent a path to improving student engagement by allowing instructors and students to interact during lectures through question and answer sessions mediated by laptops, handheld devices, or other technological devices, but many CRSs that rely on less engaging question types such as multiple-choice responses may have low or negligible impacts on learning gains (Hunsu et al., 2016). A CRS implementing self-explanation in a real-time lecture format may represent a path to substantially improving engagement and learning gains. While automated assessment of written responses is an active area of study (Gao et al., 2023; Urrutia et al., 2025; Bexte et al., 2024), it still represents a substantial computing challenge, particularly in the context of real-time assessment during lectures. Large language models (LLMs) provide a promising method of explainable automated assessment but require further evaluation with regards to their consistency and correctness. Additionally, privacy concerns such as attack vulnerabilities and data governance have arisen around the use of large private closed-weight LLMs (Li et al., 2024). Small, open-weight models can enable institutions to have total control over their implementations.

In this paper, we expand upon a previously developed explainable LLM-powered pipeline (Esiason et al., 2026a) used to generate knowledge component (KC) sets from a variety of instructor-generated questions, tag student responses with

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 248–258

Figure 1: The EXPLAINIT CRS. Instructors are able to use the ExplainIt interface to pose self-explanations questions during lecture, and students can respond to those questions in their own words using a student interface on their handheld devices or laptops. Students receive LLM-generated feedback on their answer and instructors receive a class-level summary of student performance on the question.

these KCs, and summarize topic-level classroom performance by testing a popular closed-weight model (ChatGPT GPT-4o-mini) against a relatively lightweight open-weight model (MistrailAI Ministral-3-3B-Instruct). In this work, we explore how consistently KCs are generated when the same instructor questions are used as input across multiple generations of KC sets. We also qualitatively evaluate how RAG utilizing course materials affects KC generation by examining whether or not generated KCs are within the scope of the intent of the question being posed, and whether or not the generated KCs are sufficient in capturing a "correct" response from a student.

Specifically, we seek to evaluate the following research questions:

- RQ1: Does including course materials as RAG context for these models result in the generation of higher quality KCs as measured by the scope of the generated content and the sufficiency of the generated KCs in capturing correct responses?
- RQ2: How consistent are the output of these models over multiple generation cycles as measured by the proportion of times that semantically similar KCs are generated across all generations given the same input? Does consistency vary by the type of question (e.g. factual recall, procedural knowledge, etc.)?
- RQ3: What, if any, other qualitative or quantitative patterns appear in generated KCs with regards to their quality?

By addressing these research questions, we contribute to the understanding of how these two models qualitatively behave in this novel application context, whether or not they are capable of generating useful KCs consistently, and whether or not closed- and open-weight model performance is comparable. Additionally, the human evaluation of this work establishes an expert dataset of KCs that can be used for future work on this style of system, such as fine-tuning.

2 Related Work

This work sits at the intersection of CRSs, real-time automated assessment, KC generation, and LLMs. All of these domains individually encompass broad swaths of work but the study of their intersection is relatively limited.

2.1 Classroom Response Systems

CRSs have been deployed in various forms at various grade levels in classrooms around the world to improve learning outcomes and engagement. CRSs can take many forms, from simple multiple-choice "clicker"-style CRSs to more elaborate systems that support freeform text answers (Sianturi and Hung, 2023; Tyagi and Alyammahi, 2025; Serrada-Sotil et al., 2025). These CRSs can improve student outcomes in these lecture environments in a number of ways. They can improve engagement, cognitive outcomes such as learning gains, and non-cognitive outcomes such as creating comfort in responding to classroom exercises (Serrada-Sotil et al., 2025; England et al., 2017). However, many of these

An instructor posed the following question during an undergraduate computer science lecture:
{question}
Create a list of knowledge components covered by an ideal response to this question. Your answer should consist of ONLY the list of knowledge components described in a few words each.

Figure 2: The KC generation prompt for the No RAG condition.

An instructor posed the following question during an undergraduate computer science lecture:
{question}
The following context was given for the question:
{RAGcontent}
Create a list of knowledge components covered by an ideal response to this question. Only generate knowledge components that you think are explicitly covered by this question and the provided context. Your answer should consist of ONLY the list of knowledge components described in a few words each.

Figure 3: The KC generation prompt for the RAG condition.

existing classroom response systems that are deployed during lectures are limited to formats such as multiple-choice questions that may not encourage students to think deeply about course material or may only support relatively shallow analyses of text responses that are of limited use for measuring student knowledge. A system that could solicit and analyze short self-explanation responses from students could greatly improve interactivity, measure learning in lectures, provide immediate feedback to students, and allow instructors to perform real-time instructional pivots during lectures based upon student understanding of course material. Self-explanation has been shown to be a powerful tool for improving student learning over other question formats (Chi et al., 1994; Chi and Wylie, 2014), but presents more difficulty in analysis.

2.2 Assessing Short Self-explanations with Automated Short Answer Assessment

Assessing short self-explanations in real time for such a classroom response system presents a substantial technological hurdle for several reasons. Much of the existing work on assessing student written responses investigates the effectiveness of automated assessment in homework, exams, longer essay questions, or intelligent tutoring systems intended to be used during testing or outside of the classroom (Gao et al., 2023; Urrutia et al., 2025; Chamieh et al., 2024; Bexte et al., 2024). In comparison, student explanations posed in the classroom tend to present some unique challenges. Student self-explanations in this format tend to be very short, fragmented, may use implication that can obscure semantic meaning in order to answer questions more quickly, and contain frequent errors made in the haste of responding (Esiason et al., 2026a). Assessing many responses like this in a lecture format with many students may simply not be feasible for an instructor to do themselves due to time constraints. Additionally, probabilistic assessments have some drawbacks. If we wish to automate the process of assessing this type of response during a lecture, then we must do so with strong theoretical grounding such that we minimize errors that risk misleading students or harming their confidence in themselves or the assessment tool (Kerslake et al., 2025; Chamieh et al., 2024; Stowe et al., 2024).

2.3 Knowledge Components and Generation

Knowledge components represent a promising avenue for automatically assessing student responses for the presence of key units of knowledge or understanding. KCs have been used across a number of disciplines in the past to assess responses to a number of different question formats such as multiple-choice, short answer, or coding exercises (Koedinger et al., 2012). While sets of KCs have historically been laboriously generated by experts, recent work has demonstrated the effectiveness

Model	Scope	Sufficiency
ChatGPT-4o-mini (N)	0.881	0.778
ChatGPT-4o-mini (R)	0.637	0.941
Mistral-3B-Instruct (N)	0.778	0.593
Mistral-3B-Instruct (R)	0.709	0.754

Table 1: Median scope and sufficiency scores across all five generations, by model and RAG (R) or No RAG (N) condition. Scope and sufficiency range from [0,1], with values closer to 0 being desirable for Scope and values closer to 1 desirable for Sufficiency.

of KCs for arbitrary questions using LLMs (Esiason et al., 2026a; Moon et al., 2025; Moore et al., 2024a; Wei et al., 2025). This enables an automated path to producing CRS that can automatically generate KCs for instructor-generated questions. In addition to being adaptable to a variety of input and scalable, a KC tagging approach enables transparency of assessment since the KC tags can be traced and corrected throughout the assessment process.

2.4 Large Language Models

LLMs have risen to prominence in generating knowledge components, as they have in many other text-generation tasks. LLMs have proven to be reasonably effective in this task in various contexts and question formats (Esiason et al., 2026a; Moon et al., 2025; Moore et al., 2024a; Wei et al., 2025). Models used for this purpose must be carefully selected, however. While "black box" LLMs frequently provide state-of-the-art performance on a variety of tasks, open models are rapidly approaching parity (Kaggle, 2025) and there are concerns about using black-box models for educational purposes. Some of these concerns include security issues, such as data storage compliance or attack vulnerabilities (Li et al., 2024). Other concerns include hardware and support limitations. Larger models can avoid hardware costs by being deployed remotely in the cloud, but may come with API costs and are infeasible to run locally in the event of data security concerns. On the other hand, smaller models deployed locally may incur up-front hardware costs but ensure total control over student data. As a result, exploring both large and small or open- and closed-weight models is necessary.

There are many options to choose from, but two models have been selected for this work for the sake of expedience in human assessment. OpenAI's ChatGPT-4o-mini (OpenAI, 2023, 2025) was selected as the closed-weight model for use in this experiment due to its prior use in KC generation and short assessment work (Moon et al., 2025; Moore et al., 2024b). MistralAI's Ministral-3-3B-Instruct (Liu et al., 2026) was chosen as the open-weight model for use in this experiment due to its reasonable performance in benchmarks (Kaggle, 2025) and open-weight status.

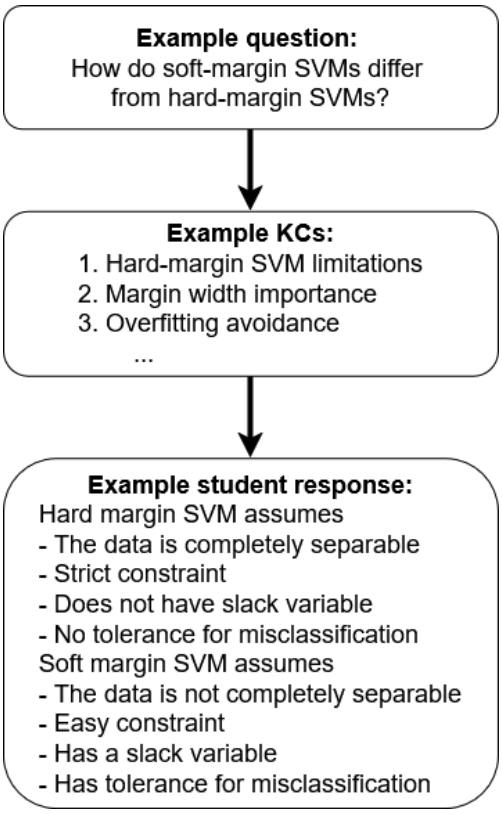

Figure 4: Example question, KCs generated from the question, and a related student self-explanation.

2.5 The ExplainIt Classroom Response System

ExplainIt is a multi-stage CRS that allows instructors to prepare short open-response questions, present them to students during lectures, gather responses from students, and provide automated feedback on those responses (Figure 1) (Carpenter et al., 2024). Instructors utilize the Instructor Dashboard to prepare questions ahead of time, and then present them to students who respond on a laptop or hand-held device of their choice through ExplainIt's web platform (Figure 4). This engages students in self-explanation, allowing for deeper learning than pure lecturing or other CRSs that may use question formats that do not promote deeper learning, such as multiple-choice. LLM-driven,

	GPT (R)	Mistral (N)	Mistral (R)
GPT (N)	18.62, $p < 0.001$	4.45, $p = 0.035$	9.88, $p = 0.002$
GPT (R)	-	5.89, $p = 0.015$	1.31, $p = 0.253$
Mistral (N)	-	-	1.21, $p = 0.272$

Table 2: McNemar's pairwise tests for differences in proportion of KCs that were out of scope by model and RAG (R) or No RAG (N) condition.

	GPT (R)	Mistral (N)	Mistral (R)
GPT (N)	12.25, $p < 0.001$	14.05, $p < 0.001$	0.08, $p = 0.775$
GPT (R)	-	39.93, $p < 0.001$	17.36, $p < 0.001$
Mistral (N)	-	-	5.97, $p = 0.015$

Table 3: McNemar's pairwise tests for differences in proportion of KCs that were sufficient for identifying correct responses by model and RAG (R) or No RAG (N) condition.

immediate, personalized feedback then aids students in troubleshooting their response. Instructors are also presented with a real-time analytical dashboard of student responses, ensuring that a human is kept in the loop of the assessment process. An important design principle of EXPLAINIT is to create a low-overhead tool for instructors to use such that integration of the tool provides benefits without adding to their workload. This has informed the automated design choices underlying this system.

Recent work with this system has assessed its ability to provide real-time instructional support in undergraduate lectures (Esiason et al., 2026a,b) (Figure 1). In this system, instructors first utilize an authoring tool to prepare questions to pose to the class, categorized by topic and stored using EXPLAINIT web platform. Then, the system uses an LLM to derive a set of relevant KCs from the question. Next, the instructor uses the Instructor Dashboard to deliver a question to students during their lecture. Students respond to this question using a web interface on their phones, laptops, or other devices. At this point, the pipeline again uses an LLM to tag student responses with the boolean presence or absence of each KC from the set generated earlier. Once all student responses are assessed, an LLM is used to generate a summary of student performance based upon the question, the student responses, and their respective KC taggings. This KC approach allows for LLM assessments to be traced from question creation to summarization, and inherently supports human-in-the-loop interfacing by allowing instructors to drill down into student responses from the summaries, all the way back to the original question posed.

Prior work has found that the system performs with a reasonable degree of accuracy with regards to agreement with human grader expectations for assessments of student responses and instructor preferences in how class-level summaries of student performance on questions should be formatted for easy and meaningful interpretation (Esiason et al., 2026a,b). While this previous work has established broad feasibility of this system, the reliability of the KC generation phase over repeated generation has not been established. The current work focuses exclusively on this stage.

3 Methods

All data collection activities were approved by IRB ANONYMIZED and were carried out in accordance with the relevant guidelines and regulations. Students were asked to review the study information sheet and electronically sign the IRB approved consent form to allow their data to be used for research purposes. Then all students were given a generic ID to interact with the system such that no personally identifiable information was included or collected by any systems external to the institution at which the research was being conducted. EXPLAINIT was deployed in one undergraduate computer science course on applied learning and data analytics in the fall semester of 2024. 27 instructor-generated questions were collected for use in this study through the Instructor Dashboard.

Models were selected to represent a balance between the state of the art in LLMs, prior work, and (in the case of Ministral-3-3B-Instruct) a relatively lightweight, open source option that could conceivably be deployed locally by an education agency. Both models were used as is out of the box. ChatGPT-4o-mini (OpenAI, 2025, 2023) was se-

Model	Median [Min,Max] KCs		
	Over (C>1)	Exact (C=1)	Under (C<1)
ChatGPT-4o-mini (N)	1 [0,4]	3 [0,5]	8 [0,16]
ChatGPT-4o-mini (R)	1 [0,3]	2 [0,6]	9 [1,13]
Mistral-3-3B-Instruct (N)	3 [0,10]	1 [0,7]	9 [3,24]
Mistral-3-3B-Instruct (R)	1 [0,4]	3 [0,7]	9 [3,25]

Table 4: Distributions of median over- and under-generated KCs per question as measured by consistency across 5 runs (C), by model and RAG (R) or No RAG (N).

Model	N Semantically Distinct KCs Generated	Median Semantically Distinct KCs per Question (Median [Min,Max])
ChatGPT-4o-mini (N)	347	12 [6,19]
ChatGPT-4o-mini (R)	325	13 [6,16]
Mistral-3-3B-Instruct (N)	374	12 [5,32]
Mistral-3-3B-Instruct (R)	395	13 [6,29]

Table 5: Total, median, minimum, and maximum number of KCs generated by model and RAG (R) or no RAG (R).

lected as a baseline based upon prior work with this system (Esiason et al., 2026a,b; Carpenter et al., 2024) and related asessment and KC generation tasks (). MistralAI's Ministral-3-3B-Instruct (Liu et al., 2026) was chosen in order to evaluate the performance of an open source, relatively lightweight model that could be reasonably be expected to run locally on institution hardware rather than relying on third-party closed-source APIs. All models were run at a low temperature (0.1) in order to align with previous work (Esiason et al., 2026a) and increase the reliability of the system. Outputs of each model were limited to 2,000 tokens, although outputs were frequently shorter.

Each model was used on the same question sets to produce knowledge components by RAG and non-RAG approaches. This was done to compare the performance of the two approaches and attempt to control for previously identified issues with this pipeline generating sets of knowledge components that were out of scope of the question being asked (Esiason et al., 2026a). We theorized that a RAG approach would narrow the generated knowledge component sets to what could be considered fairly expected of students, given that these questions are deployed after the students have only just been introduced to the material either by assigned reading or the lecture itself.

For the RAG models, one reference material per question was obtained from one of three sources. All materials produced were in text or table format, with figures excluded. Firstly, we attempted to obtain reference material from the course textbook (Tan et al., 2019) and transcribed the relevant paragraph(s) and/or tables. Occasionally, the questions referenced content that was not available in the textbook; in these cases reference materials were sourced from other comparable locations appropriate for an undergraduate student lecture (Simon et al., 2018; Boehmke). We attempted to retrieve comparably useful and relevant text for RAG across questions, but RAG material varied in length from a few sentences to several pages depending on the question.

Once RAG material were selected, the instructor questions were paired with their RAG material and ingested by the EXPLAINIT pipeline. The same prompt was used for each model, with the only variation being in the inclusion of the RAG material between RAG and No RAG conditions. In the No RAG condition, only the instructor question was passed to the LLM through the prompt (Figure 2). In the RAG condition, both the instructor question and the RAG material were passed to the LLM through the prompt (Figure 3). All models and RAG combinations were run 5 times for each input data, once with RAG utilizing course materials, and once without RAG. This resulted in 135 generated KC sets per model/RAG combination.

Two raters divided up the task of evaluating these KC sets along the metrics of scope and sufficiency. "Scope" intended to measure the whether or not a given KC was within the intent of the question being posed. For example, if the question was "what are the steps of k-fold cross validation" and one of the generated KCs involved comparing k-fold

	GPT (R)	Ministral (N)	Ministral (R)
GPT (N)	$V = 132.00,\ p = 0.577$	$V = 156.00,\ p = 0.583$	$V = 164.00,\ p = 0.562$
GPT (R)	-	$V = 103.50,\ p = 0.746$	$V = 139.00,\ p = 0.239$
Ministral (N)	-	-	$V = 133.50,\ p = 0.186$

Table 6: Wilcoxon signed-rank pairwise tests for differences in consistency of KC generation between models by RAG (R) or No RAG (N).

cross validation to other training algorithms then this KC would be out of scope. Sufficiency, on the other hand, is intended to measure whether or not a given set of KCs is sufficient for identifying a correct response from a student (regardless of whether or not it contains KCs that are out-of-scope). For example, if the question is "what are the steps of k-fold cross validation" and the LLM does not generate a KC corresponding to "divide the training data into k folds" then the generated KC set would be insufficient to identify a correct response since one of the first steps of k-fold cross validation is to divide the training data into k folds. Inter-rater reliability was computed using Gwet's AC1 (Gwet, 2002) using a stratified sample of the KCs such that each question and a respective KC set were each represented. Questions were also grouped by one expert grader using Bloom's revised taxonomy (Anderson and Krathwohl, 2001) as a guide to creating more narrowly-defined question types. Questions were grouped into the categories of procedural knowledge (e.g. recalling the steps of an algorithm), comparison (e.g. comparing two algorithms), explaining specific concepts of an algorithm (e.g. modulating the number of starting clusters in a clustering algorithm), or factual recall (e.g. recalling the definition of a term). While Bloom's revised taxonomy contains other criteria for categorizing questions, these four categories were the only ones observed in the data.

All knowledge components generated were qualitatively semantically matched to each other by an expert grader where appropriate in order to evaluate consistency of KC generation. For example, if two or more runs generated closely-related variations of the knowledge component "divide data into *k* folds" for the question "what are the steps of *k*-fold cross validation?", then these knowledge components were matched. Consistency was measured as the number of times a matched KC was generated divided by the number of runs. Ideally this value would approach 1, indicating that a KC was likely generated once per run. A value over 1 would indicate overly repetitive generation, while a value under 1 indicates inconsistent generation.

4 Results

One generated KC set was excluded from analysis of the Ministral-3-3B-Instruct model with RAG due to substantial generation failure. An exorbitant number of repetitive KCs were generated in this case (126 in total), indicating that the model may have gotten stuck in a generation loop due to the low temperature. No other models or runs demonstrated any otherwise equivalently catastrophic looping failures, although one run of one question ("what are the common properties of similarity?") for the Mistral-3-3B-Instruct produced completely topically irrelevant content in the form of software engineering advice.

IRR was achieved in the first round of coding for sufficiency (AC1=0.777). Scope required an additional round of coding on a second similarly stratified sample, at which point sufficient IRR was achieved (AC1=0.704). Once sufficient IRR was achieved for both metrics, the remaining data were divided up among the raters and coded. Differences in scope and sufficiency shown in Table 1 were measured as statistically significant by Cochran's Q test ($Q = 22.654,\ p < 0.001$ and $Q = 47.571,\ p < 0.001$, respectively). Post-hoc McNemar tests (Tables 2 and 3) demonstrate pairwise differences between models and conditions.

All models produced similar numbers of semantically distinct KCs across all runs, with the Ministral-3-3B-Instruct model producing slightly more KCs (374 without RAG, 395 with RAG) than GPT-4o-mini (347 without RAG, 325 with RAG). Both models without RAG produced a median of 12 KCs per question while both models with RAG produced a median of 13 KCs per question, although the range in number of generated KCs differed (Table 5). All models had the same overall median consistency score of 0.60. Consistency of generation for GPT-4o-mini without RAG, GPT-4o-mini with RAG, Ministral-3-3B-Instruct

without RAG, and Ministral-3-3B-Instruct with RAG ranged from [0.2,2.8], [0.2,2.4], [0.2,5.0], and [0.2,3.0], respectively. The questions that produced the most consistent KC sets (median consistency ≥ 0.80) varied by model with no readily apparent pattern. Pair-wise Wilcoxon signed rank tests were performed between model and RAG conditions (Table 6) but revealed no statistically significant difference in consistency of generation. While the dataset contains several types of questions (such as explaining the steps of an algorithm, explaining how modulating a parameter will impact algorithm performance, etc.) there appeared to be no statistical difference in the distribution of types of questions (comparison, algorithm component explanation, algorithm steps, and factual recall) between the most and least consistent questions (Fisher's Exact $p = 1.00$).

5 Discussion

With regard to RQ1 ("Does including course materials as RAG context for these models result in the generation of higher quality KCs as measured by the scope of the generated content and the sufficiency of the generated KCs in capturing correct responses?"), clear patterns can be observed in RAG vs. No RAG as well as model choice conditions. Statistical testing indicates that the inclusion of RAG content can substantially limit the generation of out of scope KCs. In GPT-4o-mini's case, out of scope KCs were reduced by 24.4%. In Ministral-3-3B-Instruct's case, out of scope KCs were reduced by 6.9%. GPT-4o-mini statistically outperforms both conditions of the Ministral model, although the difference is fairly small between the RAG conditions (0.072). The inclusion of RAG material similarly improves the sufficiency of questions in capturing correct responses, with the RAG conditions of GPT-4o-mini and Ministral showing no statistically significant difference in performance. These findings indicate that, for these metrics, GPT-4o-mini and Ministral are near or equivalent in performance and thus both potentially suitable for further exploration of this task. As a result, it may be feasible to allow institutions of higher education to maintain control over their own data using smaller, open-weight models while using a CRS like EXPLAINIT. Additionally, the findings indicate that the inclusion of course material excerpts as RAG content improve the generated KCs over the No RAG condition. Since course material would presumably be readily available to these models as it has been in this experiment, RAG presents an accessible path to improving KC generation consistency.

With regard to RQ2 ("How consistent are the output of these models over multiple generation cycles?"), consistency was moderate and there was no statistically significant difference in consistency across models and conditions. This may indicate that while issues regarding scope and sufficiency can be reduced with RAG material, issues of consistency even at very low temperatures may require other avenues of model improvement in addition to RAG such as fine-tuning, different prompting strategies such as zero-shot with instructor examples, or purpose-built models. This work may also have identified a relationship between proximity of the KC to the start of the generated KC set and consistency, although this was not explicitly tested or controlled for in this work. We observed that the most consistently generated KCs appear to generally be placed earlier in the lists of KCs produced than those of lower consistency. Because of the sequential nature of the generation of lists of KCs by LLMs in this study, we theorize that the apparent high stability of KCs that are placed first in the list may be due to these KC tokens having high probabilities given the previously-generated text and low temperature. While no statistical tests were performed on this topic, consistency scores were observed to largely decrease as the KCs moved further from the start of the KC set. This may indicate that KC sets are likely to start with high-probability tokens before steadily degenerating to lower-probability, more inconsistently generated KCs as inference proceeds. This could present an avenue for improving consistency of output by eliminating low-probability tokens at time of generation, such as by examining token log probabilities during generation.

With regard to the final RQ ("What, if any, other qualitative or quantitative patterns appear in generated KCs with regards to their quality?"), no visually obvious quantitative or qualitative patterns appeared in the generation of KCs vs. the types of questions analyzed. This may indicate that the models tested are equally as effective in generating KCs for the given dataset regardless of whether the question is oriented towards formats such as factual recitation, exploring the impact of altering or explaining individual algorithmic components, asking students to reason through the steps of an al-

gorithm, or offering a comparative analysis of two or more algorithms or methods. Such flexibility will be useful, as this would theoretically allow instructors to pose a wide range of types of questions based upon the material being delivered.

6 Limitations

Several limitations apply to this work. Firstly, the question set is varied in computer science topic and question type (factual recall, explanation of algorithm components, etc.) but still relatively small with only 27 questions. This limits statistical power when attempting to examine relationships between metrics and question type. Secondly, one catastrophic KC set generation and one completely off topic generation were observed in the open-weight model. While this may simply indicate the existence of relatively rare generation errors, these errors cannot be overlooked in an educational setting and warrant further investigation into eliminating these catastrophic generations or enhancing the human-in-the-loop aspects of the EXPLAINIT system. Additionally, RAG materials varied in quality. The size of the passages varied substantially from a few sentences or bullet points to several pages and could contain irrelevant content that could misguide the LLMs generation process, resulting in KCs that are sourced from the RAG materials but ultimately out of scope of the spirit of the question being asked. Furthermore, KCs were taken as given based upon prior work and were not validated on student data. While the expert graders evaluating the data are knowledgeable about the subject matter of the questions, a more thorough analysis of the KCs on student data would improve the robustness of the status of the generated KCs as "true" KCs. Lastly, this work failed to show within-model boosts to consistency through the use of RAG so further exploration of methods to improve consistency must be pursued. But this work highlighted a potential avenue for improving this issue by identifying a potential pattern in KC generation that could be studied and exploited in order to limit generations to KCs that the LLM is more "confident" in by restricting KCs to those with high token log probabilities. Further exploration would also inform a deeper understanding of consistency of generation, which could in turn inform how a practitioner would interpret or even tweak the KCs produced by this CRS.

7 Conclusion and Future Work

The inclusion of a pipeline allowing instructors to deploy and automatically assess self-explanation questions in CRSs has significant potential to improve learning outcomes over traditional, less interactive lecture experiences or CRSs with less engaging question types by increasing interactivity and prompting students to explain their understanding of course topics and materials in their own words. Assessing the responses to these questions accurately and consistently in real time during lectures poses a substantial technological challenge, but LLMs supported by RAG using course materials may present a promising path towards tackling this challenge through the flexible automated generation of knowledge components.

Our work addresses this novel problem by evaluating the quality of KCs generated by two popular LLMs with or without relevant RAG material drawn from course materials along several metrics: whether or not the generated KCs are within the scope of the question, whether or not the generated KCs are sufficient to identify a correct answer, and the consistency of the generation of these KCs across multiple rounds of generation. Results indicate that the LLM-generated KCs are of higher quality when relevant course material is provided for RAG. Additionally, while the closed-weight model's performance is statistically significantly better than the open-weight comparison model the gap is small. Furthermore, no statistically significant difference in consistency of KC generation is observed between RAG conditions. These results indicate that both large closed-weight and smaller open-weight models show promise in this task and that fine-tuning or improvements to the quality of RAG material may be necessary to improve consistency and quality further.

Acknowledgments

This research was supported by funding from the National Science Foundation under grants DUE-2111473, DUE-2111216, and CISE Graduate Fellowship CISE-2313998. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Additionally, the authors would like to thank Rebecca Zarch and Tyler Clark of SageFox Consulting Group for their contributions in ideation.

References

Lorin W. Anderson and David R. Krathwohl. 2001. *A Taxonomy for Learning, Teaching, and Assessing : a Revision of Bloom's Taxonomy of Educational Objectives*. Longman.

Marie Bexte, Andrea Horbach, Lena Schützler, Oliver Christ, and Torsten Zesch. 2024. Scoring with confidence? – exploring high-confidence scoring for saving manual grading effort. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 119–124, Mexico City, Mexico. Association for Computational Linguistics.

Bradley Boehmke. Regularized regression. Accessed: 2026.

Dan Carpenter, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng, and James Lester. 2024. Assessing student explanations with large language models using fine-tuning and few-shot learning. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 403–413, Mexico City, Mexico. Association for Computational Linguistics.

Imran Chamieh, Torsten Zesch, and Klaus Giebermann. 2024. 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)*, pages 309–315, Mexico City, Mexico. Association for Computational Linguistics.

Michelene T. H. Chi and Ruth Wylie. 2014. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4):219–243.

Michelene T.H. Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian Lavancher. 1994. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439–477.

Benjamin J. England, Jennifer R. Brigati, and Elisabeth E. Schussler. 2017. Student anxiety in introductory biology classrooms: Perceptions about active learning and persistence in the major. *PLOS One*, 12(8).

Jordan Esiason, Priyanka Khare, Claire Aguiar, Dan Carpenter, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng, and James Lester. 2026a. An explanation-based classroom response system for real-time analysis of undergraduate students' natural language explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(48):40822–40830.

Jordan Esiason, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng, Yeil Jeong, and James Lester. 2026b. Topic-level feedback summarization for an explanation-based classroom response system. In *Proceedings of the 57th ACM Technical Symposium on Computer Science Education V.2*, SIGCSE TS 2026, page 1325–1326, New York, NY, USA. Association for Computing Machinery.

Rujun Gao, Hillary E. Merzdorf, Saira Anwar, M. Cynthia Hipwell, and Arun Srinivasa. 2023. Automatic assessment of text-based responses in post-secondary education: A systematic review. *ArXiv*, abs/2308.16151.

Scott Grissom, Renée Mccauley, and Laurie Murphy. 2017. How student centered is the computer science classroom? a survey of college faculty. *ACM Trans. Comput. Educ.*, 18(1).

Kilem L. Gwet. 2002. Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity.

Nathaniel J. Hunsu, Olusola Adesope, and Dan James Bayly. 2016. A meta-analysis of the effects of audience response systems (clicker-based technologies) on cognition and affect. *Computers Education*, 94:102–119.

Kaggle. 2025. Mmlu open benchmarks. Accessed: 2025-03-12.

Chris Kerslake, Paul Denny, David H. Smith, Juho Leinonen, Stephen MacNeil, Andrew Luxton-Reilly, and Brett A. Becker. 2025. Exploring student reactions to llm-generated feedback on explain in plain english problems. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSETS 2025, page 575–581, New York, NY, USA. Association for Computing Machinery.

Kenneth R. Koedinger, Albert T. Corbett, and Charles Perfetti. 2012. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36(5):757–798.

Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, Bo Li, Bingsheng He, and Dawn Song. 2024. Llm-pbe: Assessing data privacy in large language models. *Proc. VLDB Endow.*, 17(11):3201–3214.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 101 others. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

Hyeongdon Moon, Richard Lee Davis, Seyed Parsa Neshaei, and Pierre Dillenbourg. 2025. Using large multimodal models to extract knowledge components for knowledge tracing from multimedia question information. In *Proceedings of the 18th International Conference on Educational Data Mining*, pages 342–353, Palermo, Italy. International Educational Data Mining Society.

Steven Moore, Robin Schmucker, Tom Mitchell, and John Stamper. 2024a. Automated generation and tagging of knowledge components from multiple-choice questions. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, page 122–133, New York, NY, USA. Association for Computing Machinery.

Steven Moore, Robin Schmucker, Tom Mitchell, and John Stamper. 2024b. Automated generation and tagging of knowledge components from multiple-choice questions. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, page 122–133, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2023. Gpt-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf. Accessed: 2025-04-23.

OpenAI. 2025. Chatgpt-4o-mini api. https://platform.openai.com/docs. Accessed: 2025-04-23.

Jaime Serrada-Sotil, Juan Antonio Huertas Martínez, and Miriam Granado-Peinado. 2025. Do audience response systems truly enhance learning and motivation in higher education? a systematic review. *Humanities and Social Sciences Communications*, 12.

Alex Dharmawan Sianturi and Ruei-Tang Hung. 2023. The challenges of using kahoot! in teaching and learning in higher education – a systematic review. In *Proceedings of the 6th International Conference on Digital Technology in Education*, ICDTE '22, page 72–77, New York, NY, USA. Association for Computing Machinery.

Laura Simon, Derek Young, and Iain Pardoe. 2018. 2.1 - what is simple linear regression? Accessed: 2026.

M. Stains, J. Harshman, M. K. Barker, S. V. Chasteen, R. Cole, S. E. DeChenne-Peters, M. K. Eagan, J. M. Esson, J. K. Knight, F. A. Laski, M. Levis-Fitzgerald, C. J. Lee, S. M. Lo, L. M. McDonnell, T. A. McKay, N. Michelotti, A. Musgrove, M. S. Palmer, K. M. Plank, and 12 others. 2018. Anatomy of stem teaching in north american universities. *Science*, 359(6383):1468–1470.

Kevin Stowe, Benny Longwill, Alyssa Francis, Tatsuya Aoyama, Debanjan Ghosh, and Swapna Somasundaran. 2024. Identifying fairness issues in automatically generated testing content. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 232–250, Mexico City, Mexico. Association for Computational Linguistics.

Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. 2019. *Introduction to Data Mining, Second Edition*. Pearson Education, Inc.

Sanjay Kumar Tyagi and Aishah Mohamed Saeed Abdulla Alyammahi. 2025. Adoption of interactive learning tools in higher education: Insights for an ai-driven future. *2025 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, pages 681–686.

Felipe Urrutia, Cristian Buc, Roberto Araya, and Valentin Barriere. 2025. Unsupervised automatic short answer grading and essay scoring: A weakly supervised explainable approach. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 38–54, Vienna, Austria. Association for Computational Linguistics.

Yumou Wei, Paulo Carvalho, and John Stamper. 2025. Kcluster: An llm-based clustering approach to knowledge component discovery. In *Proceedings of the 18th International Conference on Educational Data Mining*, pages 228–240, Palermo, Italy. International Educational Data Mining Society.