Learning Analytics for Assessment Preparation: Constructing Graphs to Guide Higher-Order Question Generation

Joel Weijia Lai
Nanyang Technological University
Singapore, Singapore
joellai@ntu.edu.sg

Shen Yong Ho
Nanyang Technological University
Singapore, Singapore
hosy@ntu.edu.sg

Fun Siong Lim
Nanyang Technological University
Singapore, Singapore
LIM_Fun_Siong@ntu.edu.sg

Abstract

High-quality assessment at scale requires questions that are conceptually rich, appropriately difficult, and cognitively engaging. We present a learning analytics pipeline that transforms archival assessments and learner performance data for question generation. Past assessment questions from an undergraduate engineering physics course are multi-labelled with topics from a predefined syllabus, and a primary large language model (LLM) is selected via inter-rater reliability against expert judgments to ensure label quality. Using these labels and student performance data, we construct an undirected weighted graph where nodes represent topics and edges capture both co-occurrence and difficulty, with student performance used as a proxy. Exploring this graph reveals difficulty-calibrated topic sets that can be retrieved around target topics. We then evaluate LLM-generated questions for difficulty and alignment with the Structure of Observed Learning Outcomes taxonomy when provided with the requested topics. Empirically, we found that GPT-5 Thinking shows higher agreement with expert labels. We further found that the generated questions, grounded by the topic-performance graph, were deemed to be more difficult and of higher-order intent under blinded expert review. The contribution is a scalable creation of higher-order assessment items with appropriate difficulty that are imperative for generative personalised learning and assessment systems.

CCS Concepts

• **Applied computing** → **Education**; **Physics**.

Keywords

Learning analytics, Performance-weighted topic co-occurrence graph, Inter-rater reliability, Higher-order question generation, SOLO relational level

ACM Reference Format:
Joel Weijia Lai, Shen Yong Ho, and Fun Siong Lim. 2026. Learning Analytics for Assessment Preparation: Constructing Graphs to Guide Higher-Order Question Generation. In *LAK26: 16th International Learning Analytics and Knowledge Conference (LAK 2026), April 27–May 01, 2026, Bergen, Norway*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3785022.3785091

This work is licensed under a Creative Commons Attribution 4.0 International License.
LAK 2026, Bergen, Norway
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2066-6/26/04
https://doi.org/10.1145/3785022.3785091

1 Introduction & Recent Works

The advancement of large language models (LLMs) has shaped various domains of education, notably the generation and automation of mass instructional content and personalised assessment tools. These models, such as OpenAI ChatGPT, Google Gemini, and their counterparts, demonstrate exceptional fluency and coherence in generating educational materials. However, recent research indicates that LLMs often underperform when producing high-quality assessment questions that require higher-order cognitive engagement [6, 10]. While LLMs are adept at generating factual or multistructural questions, which require learners to recall or list multiple discrete ideas, they struggle with designing questions that demand integration and synthesis of concepts, which are hallmarks of relational-level learning of the Structure of Observed Learning Outcomes (SOLO) taxonomy [1].

Recent efforts to address these limitations have turned to knowledge graphs (KGs) as a structural aid in question generation. KGs represent knowledge through interconnected nodes and edges, capturing the relationships between concepts in a machine-readable way [2]. When paired with LLMs, these graphs provide a scaffolding for identifying rich conceptual links thereby ensuring that questions target deeper understanding [3, 12]. For example, KAQG (Knowledge-Graph-Enhanced Question Generation) combines retrieval-augmented generation with Bloom's taxonomy and Item Response Theory to control question difficulty [4]. Similarly, KG-QAGen leverages graph structures to guide long-context comprehension tasks in LLMs, demonstrating improved coherence and topical relevance in generated questions [11]. Beyond knowledge representation, the integration of graph algorithms, such as shortest path or centrality-based traversal, allows the system to explore concept relationships dynamically, aiding in discovering non-obvious but relevant question stems. When fused with natural language prompt engineering, concept-tagging of past assessments, and fine-tuned synthesis using LLMs, this hybrid approach shows promise in moving automated assessment question generation closer to human-like quality.

Framed through the lens of learning analytics, we position relational-level question generation as both an analytic intervention for assessment design and a mechanism for adaptive learner support. A core challenge lies in engineering automated pipelines that generate conceptually congruent questions aligned with the SOLO taxonomy's relational level, while remaining sensitive to students' evolving knowledge states. With advances in generative artificial intelligence (GenAI), there is now an opportunity to dynamically produce questions that diagnose knowledge gaps and deepen understanding. Our research addresses this methodological gap by developing a learning-analytics–driven approach to question generation. Anchored in a topic–performance graph, the system delivers

assessment-ready items that are both curriculum-aligned and attuned to appropriate taxonomy levels. In doing so, we explore the role of learning analytics to proactive assessment design and adaptive learner practice.

This paper is written with the non-technical educator in mind. The goal is to present a conceptually accessible framework demonstrating how educators can work with existing question banks and student performance scores to guide AI systems toward generating meaningful assessment questions.

2 Methodology

This study constructs a topic-performance graph using topic labeling of assessment question items and data derived from student scores on these assessment questions. It captures relationships between concepts based on co-occurrence in assessments and patterns of student performance. The following subsections describe the steps to identify a suitable LLM to perform topic tagging, build a topic-performance graph, and generate questions of assessment-level difficulty at the relational level of the SOLO taxonomy. An overview of the methodology is illustrated in Figure 1. The codes and processed data produced by this work can be requested at the repository [8].

2.1 Dataset

The dataset is drawn from the formative assessments of an introductory undergraduate engineering physics course conducted in 2016 at Nanyang Technological University and consists of two midterm examinations. Midterm 1 included 11 questions: 8 short-answer questions and 3 long-form questions, with the first two long-form questions subdivided into 3 parts each, and the third into 4 parts. Midterm 2 included 13 questions: 10 short-answer questions and 3 long-form questions, each subdivided into 3 sub-questions. This gives a total of 18 and 19 sub-questions for the midterms, respectively, for a total of 37 items. A total of 1,131 students completed both midterms. The scores of all students were recorded by sub-questions and anonymised.

The structure used to organize the course content is derived from an integration of the course content, as designed by the course lecturer, and the prescribed textbook. These two sources were cross-referenced to ensure alignment between the intended learning outcomes and the actual instructional content. The content was then systematically organized into a four-tiered hierarchy, reflecting the conceptual depth of the course. The four types of entities are Unit, Chapter, Topic, and Subtopic. Units represent the broadest thematic divisions within the course, corresponding to major themes of study as outlined in the syllabus. Chapters are subdivisions within units that group related content under a coherent subtheme. Topics provide a more focused level of content organization within each chapter. They correspond to specific subject areas, instructional objectives, or sections of text. Lastly, subtopics are the most granular level of content, capturing individual concepts, procedures, or examples discussed within a topic. In total, there were four units, 10 chapters, and 74 topics and subtopics.

This hierarchical labeling system enables both human and computational interpretation of the curriculum structure. It supports clear traceability from high-level curriculum units down to individual learning concepts. As an example, the subtopic '(2.4.2.2) Inelastic Collisions' is from the topic '(2.4.2) Conservation of Momentum', which is from the chapter '(2.4) Impulse and Momentum', under the unit '(2) Mechanics'.

2.2 Topic Tagging

To establish a structured mapping between assessment content and curriculum concepts, a tagging procedure was applied to assessment questions using only the subtopic and topic levels of the content hierarchy. Each assessment question, including sub-questions, was analyzed and tagged based on the concepts it was designed to assess. This allowed for fine-grained annotation of the question content and ensured accurate alignment with the curriculum structure. Topic tags were only used if the sub-question assessed all of its subtopics, or if a topic had no subtopics. Henceforth, we address these collective tags simply as 'topics' for brevity.

The tagging process was carried out independently by one human rater with content expertise and by two state-of-the-art commercial reasoning LLMs (GPT-5 Thinking and Gemini 2.5 Pro). To ensure consistency and reliability, an inter-rater agreement analysis was conducted. This hybrid approach, combining expert human judgment with automated reasoning models, enabled a balance between interpretive accuracy and efficiency. This balance is also practical and realistic since courses are often managed by a single coordinator. The meta-prompt for topic tagging is provided in Appendix A.1.

2.3 Topic-Performance Graph Generation

While traditional knowledge graphs capture relationship between nodes in the graph, our task requires more than just these relations. Thus, we leverage the capabilities of weighted, undirected graphs to capture other information. We call this the topic-performance graph. This graph is not intended to reflect hierarchical or overt semantic relationships which are already captured in the topic list, but rather to model how curriculum concepts co-occur and interact in the context of actual assessment questions. The nodes of the topic-performance graph correspond to the same set of curriculum entities as the topic list. The edges in the graph are weighted based on co-occurrence in assessment questions and patterns in student marks. Edges are introduced between pairs of concepts that appear together in the same question part, based on the tagging described previously.

To understand how strongly two topics are connected and their associated difficulty, we consider two things: First, if two topics frequently appear together in assessment questions, it suggests they are conceptually related. Second, if students tend to score in a certain pattern on a set of topics, it gives us a proxy for the level of difficulty. We combine these two ideas into a single number, the edge weight, which tells us both how often the topics co-occur and how strongly their performances are linked. The edge weight is, thus, given by

$$w_{ij} = f_{ij} \cdot |\rho_{ij}|, \tag{1}$$

where i and j are two topics. The first term f_{ij} denotes the frequency of co-occurrence, while the second term $|\rho_{ij}|$ is the Pearson

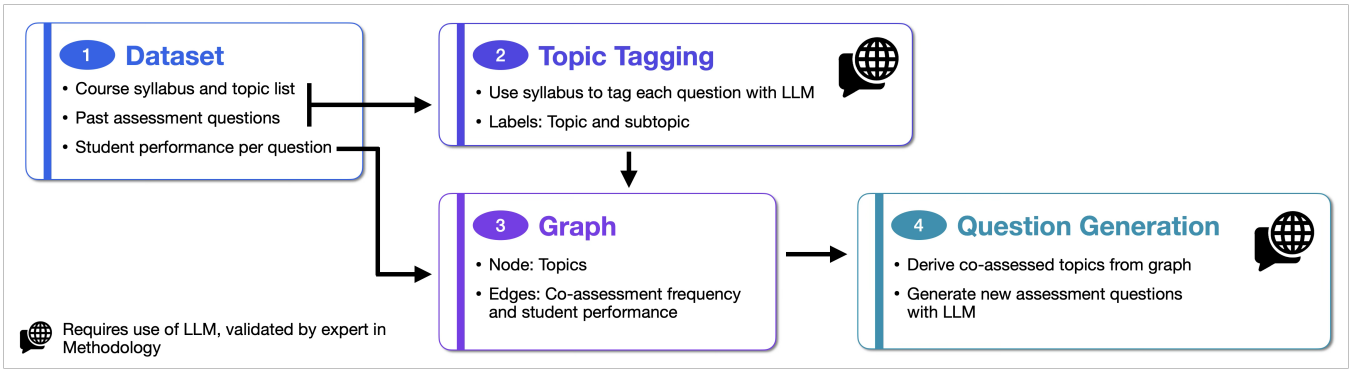

Figure 1: End-to-end methodology for building a topic-performance graph for generating new assessment questions. LLM usage is validated by a domain expert.

correlation coefficient, which denotes the strength of the performance relationship. The absolute value ensures that both positive and negative correlations contribute to the edge's magnitude. The mathematical definition of these is presented for completeness, but not discussed in detail. Let:

- Q_{ij} be the set of all sub-questions tagged with both i and j.
- s_q be the maximum possible score for sub-question $q \in Q_{ij}$.
- $s_k^{(q)}$ denote the score obtained by student k on question q, where $k = 1, 2, \ldots, N_q$, where N_q is the number of students who answered sub-question q.

Then, the co-occurrence frequency is given by $f_{ij} = \sum_{q \in Q_{ij}} 1$, which counts the number of sub-questions in which topics i and j were jointly tagged. To compute performance correlation, we first obtain normalized student scores at the subquestion level, $\hat{s}_k^{(q)} = s_k^{(q)}/s_q$, which rescales each student's score on a subquestion to the range $[0, 1]$ so that scores from questions with different maxima are comparable. Since each subquestion can be tagged with one or more topics, this normalized score is then assigned to every topic label attached to that subquestion. Aggregating over all subquestions that share the same topic label gives the topic level normalized scores $\hat{s}_k^{(i)}$. he performance correlation between two topics i and j is then given by the Pearson correlation coefficient

$$\rho_{ij} = \frac{\sum_k \left(\hat{s}_k^{(i)} - \bar{s}_i\right)\left(\hat{s}_k^{(j)} - \bar{s}_j\right)}{\sqrt{\sum_k \left(\hat{s}_k^{(i)} - \bar{s}_i\right)^2}\sqrt{\sum_k \left(\hat{s}_k^{(j)} - \bar{s}_j\right)^2}}, \tag{2}$$

where, $\hat{s}_k^{(i)}$ and $\hat{s}_k^{(j)}$ are the resulting topic level normalized scores for student k, and $\bar{s}_i$ and $\bar{s}_j$ are the their respective means. Consequently, the topic-performance graph, represented as an undirected weighted graph $G = (V, E)$, where V denotes the set of topics, and $E \subseteq V \times V$ consists of weighted edges w_{ij}.

The topic-performance graph is entirely data-driven and empirically constructed. It provides a complementary view of how the curriculum is engaged with in practice, highlighting concept associations that may not be explicitly designed but emerge through assessment patterns. By integrating topical and performance data, the resulting graph evolves from a static representation of curricular content into a dynamic, performance-aware model of learning that can be updated with more assessment items and student performance scores. The graph offers a complementary, evidence-based view of how the curriculum is enacted in practice, surfacing concept associations that were not explicitly designed but emerge through question design. By fusing topical metadata with past student performance scores, the graph evolves from a static map of content into a dynamic, performance-aware model of learning that updates as new items and results are available.

2.4 Topic Set Selection

The objective of topic set selection is to identify clusters of topics that are both empirically aligned in difficulty and structured to reflect co-occurrence patterns that support relational-level reasoning. This step bridges the empirical performance data with the cognitive framework of the SOLO taxonomy through graph-based topic retrieval. Recall, the edge weight w_{ij} is interpreted as a composite measure of empirical relatedness, where higher values indicate that the topics are not only frequently assessed together but also exhibit strong student performance patterns.

Given randomly selected start and target topics $s, t \in V$, our goal is to find topic sets that can reasonably be assessed together, to satisfy relational-level requirements, and among such sets, to choose the ones that appear more demanding, to address the difficulty requirement. To estimate difficulty, we define the difficulty score, $d_{ij} = 1/w_{ij}$. For a candidate topic set $S \supseteq \{s, t\}$, the mean empirical difficulty is computed as

$$D = \frac{1}{|S|} \sum_{(i,j) \in S} d_{ij}. \tag{3}$$

We obtain these candidate sets by performing a graph searching algorithms on G.

2.5 Question Generation Design

Lastly, to demonstrate the practical implication of our question generation framework, we conducted two trials to evaluate the effectiveness of our model, all involving the same prompt template:

- Trial 1 (T1): Two randomly chosen topics only; and
- Trial 2 (T2): Topics derived from graph exploration with advanced difficulty.

To conduct the trials, we first randomly selected two topics from the 'Mechanics' unit. These topics form the topic set for T1. Then, we used a path-finding algorithm to find all topics along the paths from one of the two randomly chosen topics to the other. We only considered candidate sets to be paths with length at most 3, one more than the average number of labels (rounded up) assigned by GPT-5 Thinking during the initial topic tagging phase. Thus, in this case, (i, j) in Equation (3) are unique edges in the path S. The topic set, defined as the candidate set with the highest D, is passed to the language model for question generation, giving us T2. It is instructive to note that such a topic set might not exist. In such cases, it could lead to the LLM hallucinating. Thus, when such a case is encountered, nothing is being passed to the LLM.

The first trial simulates a student asking a question, but the LLM lacks the grounding of a topic relation or expected assessment difficulty. The second trial is grounded in the topic-performance graph, which contains topical relations and difficulty measures. Based on our measure of mean empirical difficulty, the second trial is designed to pick topic pairs that are not frequently co-assessed and have lower student performance scores. Notice that only the topics are passed to the LLM, not the value of D.

Next, to generate questions that preserved difficulty and relational-level SOLO complexity, we designed the prompt used for question creation with GPT-5 Thinking. The prompt is provided in Appendix A.2. We repeat these steps two more times, each time with different starting topics. We then blinded the inputs and randomized the question order before consulting the course lecturer, who labeled all six questions on two metrics: difficulty (easy, moderate, advanced) and SOLO taxonomy level (unistructural, multistructural, relational).

3 Results & Discussion

The labeling task assigned to the two language models was conducted on August 10, 2025. Improvements to these language models made after the experiment date might yield different analysis results. Here, we present the results from the data collected on August 10, 2025.

3.1 Topic tagging

We analyze the multi-label annotation at the sub-question level with three raters: a human expert, GPT-5 Thinking, and Gemini 2.5 Pro. For each sub-question, each rater assigns a semicolon-separated set of labels. Labels are hierarchical and written in dot notation (e.g., 1.2.3.4), where the number of dot-delimited components defines the depth. The expert, GPT-5 Thinking, and Gemini 2.5 Pro assigned an average of 2.41, 2.45, and 3.32 labels, respectively.

We perform hierarchy-aware analysis to find Krippendorff's α at each depth, using the average Jaccard distance [7]. Our use of Jaccard distance is appropriate for unordered, multi-label sets, treating partial overlap as partial agreement. The depth-restricted analysis omits labels shorter than the target depth; it isolates exact-level consistency, suitable for preliminary analysis. To quantify uncertainty, we performed nonparametric bootstrapping by resampling items with replacement for 2,000 replicates. The results, with 95% CI, are presented in Table 1.

Table 1: Krippendorff's α by Depth with 95% CI.

Depth	α [95% CI]
1	0.624 [0.472, 0.753]
2	0.667 [0.563, 0.756]
3	0.545 [0.456, 0.633]
4	0.490 [0.362, 0.615]
Overall	0.555 [0.461, 0.630]

Next, we evaluated the pairwise reliability of GPT-5 Thinking and Gemini 2.5 Pro against the expert rater using Krippendorff's α with a Jaccard distance. The results, along with the 95% CI, after resampling with 2,000 replicates, are presented in Table 2. Restricting analysis to exact depths demonstrated that GPT-5 Thinking aligns more closely with the expert. Agreement decreases from depth 3 to 4, reflecting the expected difficulty of fine-grained choices where more specific subtopics provide more opportunities for divergence.

Table 2: Item-part pairwise Krippendorff's α at depth 3 and 4 with 95% CI.

Comparison (vs Expert)	Depth	α [95% CI]
GPT-5 Thinking	3	0.641 [0.542, 0.722]
Gemini 2.5 Pro	3	0.513 [0.403, 0.607]
GPT-5 Thinking − Gemini 2.5 Pro	3	0.128 [0.023, 0.229]
GPT-5 Thinking	4	0.507 [0.397, 0.600]
Gemini 2.5 Pro	4	0.431 [0.317, 0.529]
GPT-5 Thinking − Gemini 2.5 Pro	4	0.076 [−0.013, 0.171]

We note that at depth 3, the difference in α between (GPT-5 Thinking, Expert) and (Gemini 2.5 Pro, Expert) is positive with a 95% CI that excludes 0, indicating a robust advantage for GPT-5 Thinking over Gemini 2.5 Pro relative to the expert annotator. At depth 4, the estimated difference is still positive, but the CI includes 0, suggesting weaker evidence for a true gap at the most specific level [9]. Importantly, with this result, we selected GPT-5 Thinking's topic labels to form the topology of the topic-performance graph, also using it as the language model of choice for the question generation task.

3.2 Topic-Performance Graph

Based on the dataset and label set produced by GPT-5 Thinking, we obtain a graph that contains 30 nodes (topics and subtopics) and 53 edges. Figure 2 shows the topic–performance graph where the node size encodes the sum of incident edge weights and node color shows the community the node belongs to. The community detection algorithm used in this case is the greedy modularity algorithm. There are six isolated nodes and six communities [5]. The edge thickness encodes w_{ij}, with thicker edges representing higher weights.

3.3 Question Generation from Topic Selection

From our analysis, edge difficulty scores are approximately normally distributed with mean 1.01 and standard deviation 0.566.

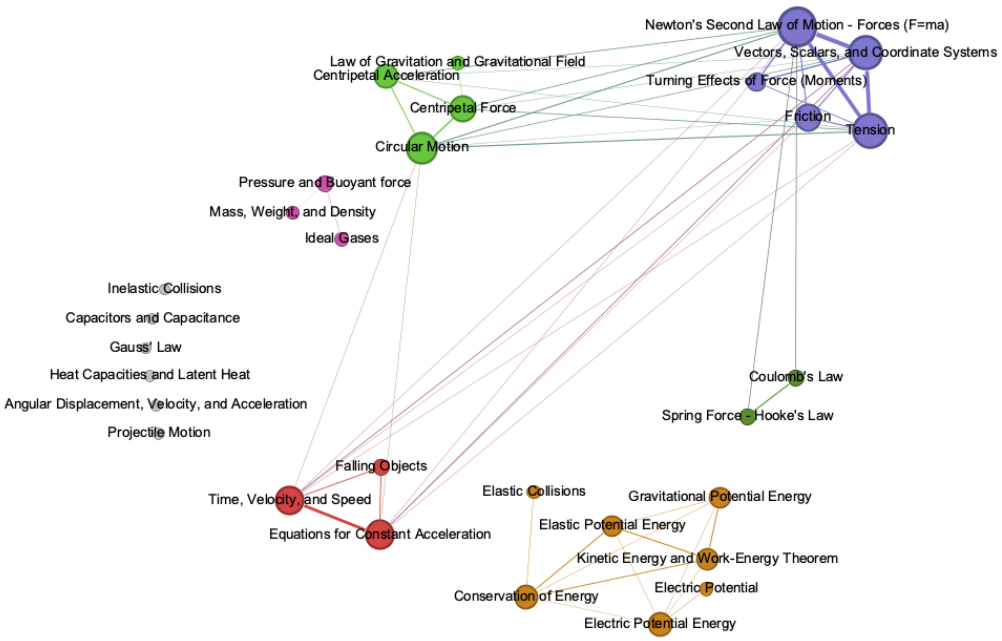

Figure 2: Visualization of the topic-performance graph. Nodes, representing topics, are sized by strength and colored by community. Edges link co-tagged topics with edge width encoding w_{ij}. We identified six isolates and six communities.

We classify a question as “Advanced" if its empirical difficulty is more than 0.5 standard deviations above the mean. Using a 0.5σ cutoff corresponds to labeling roughly the top 30% of questions as “Advanced". Table 3 lists the randomly selected starting topics and the additional topics inferred from our topic–performance graph. Questions were generated from the starting topics in both trials; in T2 we also used the additional topics. In total, six questions were produced, one for each topic set and presented in random order, the course lecturer labeled all questions.

These labels reveal that T2 consistently led to more challenging and cognitively demanding questions, provided the starting topics were not already advanced. In terms of difficulty, we note that the metaprompt does not explicitly mention difficulty. Yet, when tasked with generating questions, T1 tended to produce easier questions, whereas T2 pushed them toward Advanced difficulty. In terms of cognitive engagement, T1 generally stayed at the Multistructural level, where students only needed to handle separate ideas. By contrast, T2 frequently elevated questions to the Relational level, requiring students to integrate and connect concepts into a coherent whole. This shows that through prompt engineering alone, the language model may not be able to produce Relational-level questions. However, by anchoring the subset of topics to a topic-performance graph-based methodology, it can systematically increase both the complexity and cognitive depth of the generated questions. The result is consistent with the outcome when the course lecturer was tasked with considering both the difficulty and cognitive engagement (SOLO) collectively and ranking the questions from 1 (least advanced and lowest cognitive engagement) to 6 (most advanced and highest cognitive engagement).

Lastly, we note that the difficulty value D was not provided to the LLM during question generation. Nonetheless, the evaluation indicates a relationship in which higher values of D are associated with more advanced questions. While this relationship was not examined in depth in the present work, it suggests a promising avenue for future study. Notably, when the initial topic set is characterized by low difficulty (low D), GPT-5 Thinking tends to generate questions of comparable difficulty, even without explicit information about the target level. In such cases, the resulting SOLO levels may fall short of the desired Relational stage. By incorporating historical performance data, however, we can construct topic sets that elicit more challenging questions and reach higher SOLO levels.

4 Future Work & Conclusion

The results obtained provide a prospect for the use of this methodology. Including a larger dataset will expand coverage by incorporating sub-questions and additional cohorts. Including assessment materials from multiple years will also enable analyses of temporal drift in item design and shifts in topic co-occurrence. Additionally, in our graph construction, we add an edge only when a topic pair co-occurs in at least one sub-question. Consequently, excluding isolates from graph-based metric, such as D introduces a bias that favors frequently co-occurring topics. To mitigate this, future iterations will augment edges using prior curricular links and text-based similarities, reducing isolation without inflating spurious ties. Finally, our metrics, $w_{i,j}$ and d_{ij} are scale-dependent and may not align exactly with pedagogically meaningful difficulty levels. We can decouple co-occurence and performance, so that we can also examine frequently co-assessed topics and with lower student scores.

Table 3: Topics selected randomly (Trial 1), and through graph exploration (Trial 2) alongside the resulting evaluation by the course lecturer on the generated questions.

T	Starting topics (Randomly chosen)	Additional Topics (Graph exploration)	D	Evaluated		
				Difficulty	SOLO Level	Rank
1	Kinetic Energy and Work-Energy Theorem Gravitational Potential Energy	–	0.58	Easy	Multistructural	1
2	Kinetic Energy and Work-Energy Theorem Gravitational Potential Energy	Electric Potential Energy Conservation of Energy	1.34	Advanced	Relational	5
1	Tension Circular Motion	–	0.49	Easy	Multistructural	2
2	Tension Circular Motion	Time, Velocity, and Speed Friction	1.32	Advanced	Relational	6
1	Equations for Constant Acceleration Newton's Second Law of Motion - Forces (F=ma)	–	1.74	Medium	Relational†	4
2	Equations for Constant Acceleration Newton's Second Law of Motion - Forces (F=ma)	Friction Time, Velocity, and Speed	1.34	Medium	Relational	3

† The lecturer noted that this straddles between multistructural and relational. In the table, we present the higher of the two.

We also plan to calibrate D during item generation and evaluation by employing one-/few-shot prompting with expert-annotated exemplars, and using these anchors to validate or rescale graph-derived D to levels interpretable by instructors and assessment designers.

The course lecturer observed during evaluation that the language model occasionally uses non-standard phrasing, which could be clarified through supporting diagrams. This points to multi-modal question generation as a valuable direction for future work. Furthermore, like many GenAI approaches, the method is stochastic and may yield different results across trials. Additional experiments, such as investigating correlations between the number of included topics and question complexity, would strengthen the findings and enhance the methodology.

In conclusion, this work develops a hybrid methodology that integrates graph-based learning analytics with LLMs to improve the quality of AI-generated assessment questions. This hybrid approach aligns intended curriculum structure with observed learning outcomes, enabling the creation of assessment items that better target higher-order cognitive skills. Our results show that GPT-5 Thinking is better aligned with the human rater with $\alpha = 0.641$ [0.542, 0.722] and $\alpha = 0.507$ [0.397, 0.600] at depth 3 and 4, respectively. Using systematic prompt engineering, it can generate questions that are accurately rated for cognitive complexity and engagement.

Beyond these findings, the contribution of this work lies in demonstrating how graph-based representations of curriculum knowledge and context-specific performance data can be systematically combined with GenAI to inform assessment practice. By aligning intended curricular structures with observed patterns of student performance, the framework provides a data-driven way to surface meaningful concept relationships and translate them into assessment questions. This methodological integration not only advances analytic approaches but also creates practical opportunities to close the loop with educational stakeholders. For instructors, it offers insights into curriculum alignment and tools for designing higher-order assessment questions that reflect both content and learner performance. For learners, it creates pathways to more targeted and cognitively engaging questions that can support deeper understanding and self-regulated learning.

References

[1] John B. Biggs and Kevin F. Collis. 1982. *Origin and Description of the SOLO Taxonomy.* Elsevier, New York, USA, 17–31. doi:10.1016/b978-0-12-097552-5.50007-7

[2] Tomaž Bratanič. 2024. Knowledge graphs & LLMs: Multi-hop question answering. https://neo4j.com/blog/developer/knowledge-graphs-llms-multi-hop-question-answering/

[3] Tuan Bui, Oanh Tran, Phuong Nguyen, Bao Ho, Long Nguyen, Thang Bui, and Tho Quan. 2024. Cross-Data Knowledge Graph Construction for LLM-enabled Educational Question-Answering System: A Case Study at HCMUT. In *Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia (ICMR '24).* ACM, New York, USA, 36–43. doi:10.1145/3643479.3662055

[4] Ching Han Chen and Ming Fang Shiu. 2025. KAQG: A Knowledge-Graph-Enhanced RAG for Difficulty-Controlled Question Generation. doi:10.48550/ARXIV.2505.07618

[5] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Physical Review E* 70, 6 (Dec. 2004), 066111. doi:10.1103/physreve.70.066111

[6] Nghia Duong-Trung, Xia Wang, and Miloš Kravčík. 2024. *BloomLLM: Large Language Models Based Question Generation Combining Supervised Fine-Tuning and Bloom's Taxonomy.* Springer Nature Switzerland, Krems, Austria, 93–98. doi:10.1007/978-3-031-72312-4_11

[7] Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures* 1, 1 (April 2007), 77–89. doi:10.1080/19312450709336664

[8] joellaiwj. 2025. topic-performance-graph-question-generation. https://github.com/joellaiwj/topic-performance-graph-question-generation

[9] J. Neyman. 1937. Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 236, 767 (Aug. 1937), 333–380. doi:10.1098/rsta.1937.0005

[10] Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. 2024. *Automated Educational Question Generation at Different Bloom's Skill Levels Using Large Language Models: Strategies and Evaluation.* Springer Nature Switzerland, Cham, Switzerland, 165–179. doi:10.1007/978-3-031-64299-9_12

[11] Nikita Tatarinov, Vidhyakshaya Kannan, Haricharana Srinivasa, Arnav Raj, Harpreet Singh Anand, Varun Singh, Aditya Luthra, Ravij Lade, Agam Shah, and Sudheer Chava. 2025. KG-QAGen: A Knowledge-Graph-Based Framework for Systematic Question Generation and Long-Context LLM Evaluation. doi:10.48550/ARXIV.2505.12495

[12] Runhao Zhao, Jiuyang Tang, Weixin Zeng, Ziyang Chen, and Xiang Zhao. 2024. Zero-shot Knowledge Graph Question Generation via Multi-agent LLMs and Small Models Synthesis. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24)*. ACM, New York, USA, 3341–3351. doi:10.1145/3627673.3679805

A Prompt Templates

A.1 Topic Tagging

```
You are an intelligent assistant designed to assign topic codes to user-provided questions based on a reference list from an Excel file.

The assistant receives an Excel file from the user that contains at least two key columns:
- Topic: Keyword of a topic.
- Topic Code: A unique numeric code associated with the topic.

The assistant will be provided with a user question and must:
- Analyze the meaning and intent of the question.
- Compare the question to the list of topics in the Excel file.
- Use semantic understanding to identify all relevant topics.
- Do not tag a topic simply because of keyword matching.
- Extract and return the Topic Codes that match the question, using the correct output format.

Response format:
- Output only the topic codes, separated by a semicolon followed by a space (; ), with no extra text or explanation.
- If no relevant topics are found, return "0".
- Do not answer the question, just output the topic codes.

Constraints:
- Do not include topic names in the response.
- Do not return explanations, summaries, or additional commentary.
- Always match semantically, not exact word matches.
- Topic codes must be taken directly from the Excel file.

Clarifications:
- Only include the topic code if the topic is vital in solving the question.
- Assign third-level topic codes (e.g., 1.1.5) only if **all** of its fourth-level subtopics (e.g., all 1.1.5.1, 1.1.5.2, 1.1.5.3) apply, or if it has **no** fourth-level subtopics.

Personality and Tone:
- The assistant should be efficient, neutral, and strictly focused on performing the task with precision.
```

A.2 Question Generation

```
You are an assessment question generator for an **introductory** undergraduate engineering physics course. The questions you create should be open-ended exam questions. All questions must be at the Relational level of the SOLO taxonomy, requiring students to connect and integrate the topics.

The question should:
- Be word problems describing a physical scenario with real-world applications.
- Not require complex mathematical procedures such as calculus or linear algebra.

When given a list of topics, only use those topics to generate a complete question. The question must always have a solution. Do not provide commentary or extra text.

Questions should require a student to demonstrate ability to integrate multiple ideas and recognize the relationships between them, creating a coherent, holistic understanding of a subject.
```