From concepts to learning objectives to questions: a pedagogically-aware prompting pipeline to generate questions from textbooks

Sergey Sosnovsky[1,*,†], Keyang He[2,**,‡] and Andre Bekema[3,***,§]

[1]Utrecht University, Princetonplein 5, Utrecht, 3584 CC, Netherlands
[2]Utrecht University, Princetonplein 5, Utrecht, 3584 CC, Netherlands
[3]Edu-Suite, Abe Lenstra Boulevard 50-10, 8448 JB, Heerenveen, Netherlands

Abstract

Large Language Models (LLMs) are increasingly used in AI in Education (AIED) research. However, their lack of pedagogical understanding, susceptibility to hallucinations, and limited reliability in multi-step reasoning pose challenges for effective educational use. This paper explores externally orchestrated prompt chaining enriched with domain and pedagogical information as a mitigation strategy. We apply this approach to automatic question generation from textbooks, proposing a multi-step pipeline that extracts key concepts and learning objectives to guide generated questions. The results of expert-based evaluation experiment demonstrate that the proposed approach significantly outperforms alternative approaches that do not stress pedagogical alignment or do not follow the prompt chaining strategy.

Keywords

Question Generation, Large Language Models, Generarive AI, Prompting Strategy, Textbooks, Bloom's Taxonomy

1. Introduction

Large Language Models (LLMs) have become a versatile tool in the arsenal of AI in Education (AIED) researchers, applied to a wide range of problems that require scalable language processing and generation [1]. From detecting students' errors [2] to modeling their performance [3], from producing learning content [4], to providing feedback [5], LLMs have proven their ability to automate a large set of tasks that previously required manual authoring or purposefully-developed symbolic models. Full-fledged applications employing LLM-driven tutoring approaches are also being developed by both the research community [6, 7] and industry (e.g., Khanmigo[1], Learn Your Way[2]).

Nevertheless, an over-reliance on LLMs can become counter-productive, as their decision making mechanisms and priorities are often in conflict with didactic objectives of what can be considered an "effective AIED application". LLMs lack pedagogical understanding as they are optimized for next-token prediction and not learning outcomes. As a result, LLMs cannot reason about how students learn, but only about the likelihood of user's satisfaction by the provided explanation [8]. LLMs are known to fabricate information that is linguistically plausible, but factually or semantically incorrect [9]. LLMs can struggle with multi-step reasoning, as their reasoning is statistically approximated rather than logically derived, hence small mistakes can accumulate [10]. This is exacerbated by the possibility of hallucinations and the fact that it is very hard to analyze the internal structure of an LLM and explain its results.

A possible mitigation strategy to address these limitations of LLM-based AIED methods is the use of externally orchestrated prompt chaining enriched with domain and pedagogical logic. Prompt chaining has emerged as an effective strategy to offset LLMs' drawbacks in intelligent software by structuring complex tasks into verifiable intermediate steps and reducing the risk of ungrounded or

iTextbooks'2026: Seventh Workshop on Intelligent Textbooks, June 28, 2026, Seoul, Republic of Korea
s.a.sosnovsky@uu.nl (S. Sosnovsky); keyang23h@hotmail.com (K. He); andrebekema@edu-suite.com (A. Bekema)
0000-0001-8023-1770 (S. Sosnovsky); 0009-0009-8404-6148 (K. He)

[1]https://www.khanmigo.ai/
[2]https://learnyourway.withgoogle.com/

cascading errors [11]. Unlike single prompt solutions, this approach forces step decomposition outside the model, making intermediate reasoning stages explicit rather than internally inferred. By structuring the interaction as a sequence of narrowly focused steps, each stage can be independently verified, modified, and enriched with external knowledge and/or pedagogical constraints. While this strategy does not guarantee correctness, since individual steps may still contain errors, it reduces the likelihood of uncontrolled error cascades by simplifying LLM tasks, enriching the prompts at each step with relevant pedagogical knowledge and enabling validation between steps.

This paper describes an attempt to apply this strategy to the task of question generation (QG) from instructional texts. The overarching motivation for the project has been to enable authors/publishers of textbooks to automatically enrich them with interactive exercises. This could potentially serve two purposes: (1) providing students with an opportunity to practice knowledge they have just studied and (2) offering adaptive support that would guide students through the textbook taking into account their current mastery. We propose a pipeline that breaks the overall process of QG into sub-tasks, where each sub-task is solved by a dedicated prompt. To make sure each question addresses a cohesive knowledge unit covered by the relevant section, the first step tasks the LLM to extract concepts from its text. To focus questions on the core concepts introduced or elaborated in each section, the second step weights concepts across all sections of the textbook taking into account relative importance of concepts to sections and potential prerequisite-outcome relations. To generate questions that address target concepts on the right cognitive levels, the third step prompts the LLM to extract concept-related learning objectives (LO) of the section. Finally, based on the set of weighted LOs, questions are generated. On each step, a list of instructions is included in the prompts to reduce the possibility of hallucination, ground the LLM in the content of target section of textbook and stress the pedagogical context.

The main research question of this study can be formulated as follows:
RQ: Does organizing LLM-based question generation as a chain of pedagogically aligned prompts improve the pedagogical quality of generated questions, as assessed by domain experts?
In particular, we are interested in whether the proposed strategy can outperform the currently prevalent monolithic prompting techniques [12]. To address this question, we implemented three alternative question generation strategies that reuse the same underlying prompts but differ in how these prompts are organized. We then conducted an expert-based evaluation of the generated questions across several quality metrics.

All supplementary material produced during this study including all LLM prompts, and experimental instruments are available in an online repository[1].

The rest of the paper is structures as follows. Section 2 reflects on some of the related work that has been done in the field of QG and LLM-based AIED. Section 3 presents the details of the developed multi-step QG pipeline. Section 4 describes the design of the expert-based experiment conducted to evaluate the efficacy of the four alternative QG methods. Section 5 analyses the results of the experiment. Finally, Section 6 concludes the paper by summarizing its main outcomes and limitations, and by outlining several directions for future work.

2. Related Work

Previous studies group AQG approaches by application (standalone, visual, and conversational) and distinguishes rule-based methods from neural models that learn to generate questions from data [13]. With the rapid adoption of large language models, recent AQG work increasingly uses commercial LLMs as the main generator, so we focus on GenAI and LLM-based methods for producing assessment questions from instructional text.

A common baseline in this line of work is single prompt generation, where a passage is given as context and the model produces complete questions in one call. This pattern appears in MCQ generation with a fixed prompt template by using GPT 3.5 [14] in the domain of history. Lee et al. report a teacher-facing workflow in English education where instructors choose a reading passage and a question type

[1](https://anonymous.4open.science/r/Paper-C425/)

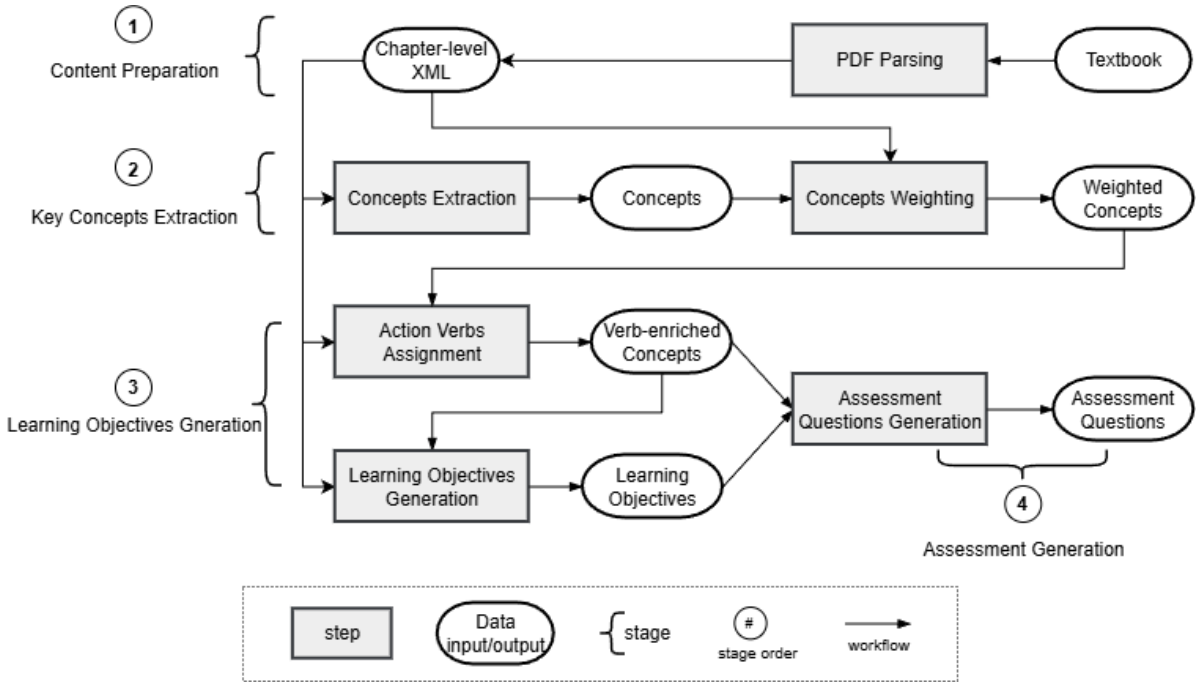

Figure 1: Overview of the pipeline

prompt, and the question is also generated in one call [15]. Chan et al. similarly refine prompt wording through iterative trials, but still generate each question directly from the given passage, sometimes with an answer included [16]. Beyond this baseline, recent work mainly adds stronger control knowledge or intermediate structures. Maity et al. prompt LLMs to generate either a set of questions for a textbook context or a six-question set aligned with Bloom's taxonomy levels in one call, showing how taxonomic prompts can steer cognitive demand [17]. Other studies target specific assessment formats and substeps, such as multiple choice cloze vocabulary items, where GPT 3.5 is used to create a masked sentence and to propose and filter distractors [18]. It has also been applied to STEM calculation items, where prompts ask for a question together with an answer scheme, and can include a verification step that uses code execution to check the mark scheme [19]. A complementary direction integrates external structure for better grounding, for example concept-guided generation with a fine-tuned T5 model and concept coverage based scoring [20], or graph-supported generation where an LLM builds and queries a knowledge graph and then generates competency questions from retrieved evidence [21].

These studies motivate two issues that our work directly targets. First, educational sources are often long and dense, and even when modern LLMs can fit a whole section in the prompt, the model still needs guidance to select suitable question targets [22]. Second, many LLM based methods either rely on a single prompt baseline or control generation mainly through surface prompt patterns, while fewer approaches provide explicit pedagogical targets that connect what to assess with how to ask. Our approach addresses this gap by using extracted concepts and LOs as explicit generation targets, so that questions are grounded in the section content but guided by instructional intent.

3. Question Generation Pipeline

3.1. Content Preparation

The overview of the proposed pipeline is presented by Fig. 1. Before engaging the LLM, we need to process the original PDF-formatted textbook. This phase is taken care of by the InTextbooks tool, which parses PDF textbooks into structured XML-based[2] representations [23]. The resulting model

[2]InTexbooks represents textbook models using the Text Encoding Initiative (TEI) format (https://tei-c.org)

separates table of contents, index, and other auxiliary parts, and dissects the main part of a textbook into chapters and sections. For the purposes of this project, we do not employ any of the additional services that InTextbooks provides, and only export an ordered list of XML files containing textual content of the sections.

3.2. Prompt structure

All LLM-based steps in the pipeline use a shared prompt template. Each sub-task is expressed through a combination of a *system prompt* that defines stable instructions, and a repeatable *task prompt* that injects task-specific inputs (e.g., for a concept extraction step, it will contain the text of a section, from which the concepts are extracted). Although concrete tasks differ, prompts follow the same structure recommended in the OpenAI's prompting guides [24, 25]. Splitting each task into two prompts improves reusability and reliability. The system prompt defines common rules across all sections and runs (role definition, execution constraints, and output schema), while the task prompt acts as the container for variable inputs. This design reduces prompt duplication, limits unintended instruction drift, and allows the same prompt logic to be applied consistently across all textbook sections. All prompts except the last one share one common component: the role/persona assumed by the LLM is a senior instructional designer. For the last prompt, LLM is instructed to act as an assessment item writer. Other details of the prompts are described in the corresponding subsections.

3.3. Concept Extraction and Weighting

Concept extraction. After the content pre-processing, the concepts are extracted from an entire textbook at the *section level.* This keeps each LLM call within practical content window limits and aligns extracted concepts with the instructional focus of corresponding sections. For each section, we pass the section heading and section text to the LLM and ask it to produce a list of instructional concepts with short, context-grounded definitions. The prompting procedure follows the shared system–task structure described in Section 3.2. The system prompt defines the role, the objective (extract key instructional concepts and definitions), and a set of instructions. The instructions are designed to guide the LLM towards extracting pedagogically important concepts and constrain its tendency to invent/infer concepts that are not present in the section.

Concept weighting. The lists of concepts extracted by LLMs are exhaustive and often include auxiliary notions despite explicit instructions to focus on "key concepts". Therefore, before passing the concepts to the next step, we ask the LLM to estimate the importance of each extracted concept *within a section.* For each section, the weighting task prompt includes the original section text and the concept list for that section. The system prompt's objective is to assign a weight [0.0; 1.0] to each concept grounded in the section context. To guide the weighting process towards defining the pedagogical importance of a concept, the prompt includes instructions stipulating the concept's centrality to the main instructional goal of the section, its relevance for knowledge assessment, etc.

3.4. Learning Objectives Identification

Knowing a list of concepts covered by a textbook section may not be enough to validly assess students' understanding of it. A section can merely introduce a concept focusing on its definition, it can underline its relations to other concepts, it can illustrate it with examples, or further explain how it is applied during problem solving. To develop assessment that accurately measures the results of instruction it should focus on the LOs of this instruction [26]. A LO is "the specific knowledge that a learner has to acquire about a concept or skill and the tasks to be performed" [27]. There exist various frameworks for formulating LOs. We have employed a widely used Bloom's taxonomy of educational objectives, in particular, its cognitive domain [28]. It organizes cognitive processes into six levels (Remember, Understand, Apply, Analyze, Evaluate, and Create) and is often used to guide the choice of action verbs that express intended learning outcomes [29]. In the proposed pipeline, we used this idea to formulate LOs of textbook sections based on the key extracted concepts.

Verb assignment. On this step, we included the action verb bank based on the Bloom's taxonomy in the system prompt and asked the model to select up to three verbs per concept that were best supported by the section text. Allowing up to three verbs captured cases where a concept supported multiple cognitive operations (e.g., identify vs. explain) and reduced potential instability due to forcing a single verb. Verb assignment was applied only to the concepts with weights above the global threshold W = 0.5. The task prompt provided the concept name and its definition together with the content of the section to allow the model to ground the selection of action verbs in levels of concept coverage by the section.

LO generation. Next, we prompted the LLM to generate LOs from the key concepts, their definitions, assigned action verbs, and the full section texts. We included a predefined verb-to-level mapping in the system prompt to group the assigned verbs by Bloom level. The set of instructions stipulated the template for wording LOs (according to [29]), constrain the number of LOs per concept-level pair, and forbade the LLM to use any sources outside of provided context.

3.5. Question Generation

QG was the final stage of the pipeline. The input for the question generation prompts included all the information accumulated at the previous steps: a list of key concepts with definitions, a list of LOs linked to corresponding concepts, text of a section. The instructions followed the previously described strategy of stressing pedagogical aspects and using the definition of the associated concept and the section content as supporting context. To ensure consistent output across sections, we enforced a uniform assessment format: for each target LO, the model generated a multiple-choice question and a fill-in-the-blank question.

4. Experiment

To evaluate the effectiveness of the proposed approach, we organized expert-based experiment on the quality of generated questions. A group of experts rated a selection of questions produced by four different LLM-based methods (including the proposed one) from the same source.

In all experiments, across all prompts and conditions, we used the GPT-5 snapshot 2025-08-07. This is a state-of-the-art model that supports long inputs and structured outputs required by the pipeline. The `seed` parameter was set to `default` and the `reasoning_effort` to `medium` to keep output quality and efficiency relatively consistent.

4.1. Four Conditions

Three alternative prompting strategies for QG were implemented alongside the proposed pipeline. The prompts' structure and all prompts' components remained unchanged across conditions. However, the alternative strategies either included fewer steps (hence implemented less pedagogical alignment) or did not employ the prompt chaining method (hence could not externalize the outcomes of intermediate steps and enforce grounding at every step).

Strategy 1: *Content* (baseline). This is a single-prompt method that takes as input section texts and generates assessment questions directly from it. It uses only system and task prompts described in Section 3.5. This strategy represents a common baseline in LLM-based QG, reflecting a class of naïve yet widely used approaches that depend on monolithic prompting.

Strategy 2: *Concepts*. This method utilizes the concepts extracted and weighted at Step 2 of the pipeline. The input for the QG prompt includes text of a section, a list of key concepts (names with short definitions) extracted from the section. The concept list specifies *what to assess*, while the section text provides the supporting context. By externalizing concept selection and making it explicit in the prompt, this strategy introduces a lightweight knowledge engineering step, which is characteristic of traditional knowledge-based adaptive approaches.

Strategy 3: *LOs* (proposed). This is the experimental condition representing the QG pipeline described in Section 3.

Strategy 4: *SuperPrompt*. This strategy implements a single-prompt approach inspired by chain-of-thought prompting. As in the ***Content*** baseline, the LLM receives only the section text as input. However, the prompt explicitly instructs the model to internally follow the logic of the proposed pipeline by first inferring key concepts, then deriving learning objectives, and finally generating assessment questions. All task-specific instructions and constraints are embedded within a single prompt, with intermediate reasoning remaining implicit and unobservable. This strategy evaluates whether a monolithic, chain-of-thought-style prompt can substitute for the explicit step-by-step logic of the proposed pipeline.

4.2. Corpus

An openly available digital textbook "*Introductory Statistics*" from OpenStax.org [30] was used as the source content. Two sections were selected for evaluation: a more introductory Section 2.5 (*Measures of the Center of the Data*) and the more advanced Section 13.2 (*The F Distribution and the F-Ratio*). Sections with differing levels of conceptual depth and complexity were chosen to safeguard against potential confounding effects related to content difficulty.

Across the sections, the four conditions resulted in different numbers of generated questions. For Section 2.5, ***LOs*** produced 26 questions, ***Concepts*** – 20, ***Content*** – 16, and ***SuperPrompt*** – 14. For Section 13.2, the counts were 30 for ***LOs***, 26 for ***Concepts***, 14 for ***Content***, and 12 for ***SuperPrompt***.

The final selection of questions given to the experts for evaluation was a balanced set of 80 items in total, with 10 randomly selected items per condition in each section [3].

4.3. Question Quality Criteria

Each item was evaluated via a structured survey covering seven quality criteria plus *Difficulty* and *Cognitive Level*. Following established practices in educational QG [31, 32], the quality criteria included:

- Grammaticality: the item is grammatically correct;
- Relevance: knowledge assessed by the item is relevant to the section;
- Answerability: only one correct answer is possible for the item;
- Interestingness: the item is interesting;
- Importance: the topic assessed by the item is important for the section;
- Utility: the item should help students learn the material of the section;
- Sufficiency: the section is sufficient to answer the item.

For these seven quality criteria, experts rated their agreement on a 5-point Likert scale (from Strongly disagree to Strongly agree). For *Difficulty*, experts used a 5-level scale rating the item from (Much too easy to Much too hard). For *Cognitive Level*, experts were asked to compare the level of knowledge addressed by the item with the level of the material provided by the section and select one of four categories: Lower, Comparable, Higher (beyond the provided material), or Too high (not answerable based on the section).

4.4. Data Collection

Eight experts participated in the study. Four experts evaluated both sections (80 assessment items in total), two experts evaluated only Section 2.5 and the remaining two experts evaluated only Section 13.2. Of the eight experts, five were professionals working on mathematics-related e-course content, one was a researcher, and two were MSc students who had teaching experience in Introductory Statistics. The sets of questions per section were sampled once, but each expert received them in a uniquely randomized order.

[3]The exact sets of questions together with the content of relevant sections can be found in the online repository

Table 1
Global likelihood-ratio (LR) tests for the fixed effect of *Method* in the LMM.

Criterion	LR statistic	df	p_{LR}
Answerability	10.4	3	.023
Relevance	6.91	3	.075
Interestingness	6.91	3	.058
Utility	5.88	3	.118
Importance	5.19	3	.103

5. Results

5.1. Data pre-processing

With 6 expert ratings per item across 20 items, we obtained $N = 120$ ratings per condition and per criterion or 480 ratings per criterion overall, except for a single missing value for the ***Content***'s *Interestingness* criterion ($N = 119$). Further quality-control analysis identified seven expert-criterion pairs with abnormal rating patterns, mainly due to straight-lining and low variance. Consequently, we had to filter out 440 individual ratings from further analysis[4].

Because the expert ratings followed an incomplete design (some experts rated 80 items while others rated 40), we estimated reliability for the numeric quality criteria using a linear-mixed-model-based intra-class correlation (LMM-based ICC) [33]. *Answerability* showed the highest agreement ($ICC_{\text{mean}} = .81$, good). *Relevance* (.55), *Interestingness* (.53), *Importance* (.56), and *Utility* (.54) were in the moderate range. *Sufficiency* showed poor reliability (.24), suggesting substantial disagreement between experts on whether the section contained enough information to answer the questions. *Grammaticality* had $ICC_{\text{mean}} = .00$, which reflects a ceiling effect where almost all items received near-maximal scores. Consequently, we removed Grammaticality and Sufficiency from further analysis.

5.2. Quality of Generated Questions

This section reports results for the remaining five Likert-style quality criteria: *Relevance*, *Answerability*, *Interestingness*, *Importance*, and *Utility*.

For each criterion, we fitted a linear mixed-effects model with *Method* and *Section* as fixed effects and random intercepts for *ItemID* and *ExpertID*. We then tested the global effect of *Method* using a likelihood-ratio (LR) test that compared the full model against a reduced model without *Method* (df = 3).

Table 1 shows that *Answerability* had a significant global method effect. *Interestingness* and *Relevance* showed non-significant trends, while *Utility*, and *Importance* did not show evidence of differences between methods.

Table 2 reports the model-adjusted marginal means (M) and the standard deviations (SD) for each method. Across all quality criteria, the experimental condition (***LOs***) consistently achieved the highest scores. This indicates a stable performance advantage of the proposed pipeline across all quality dimensions. Interestingly, none of the other strategies demonstrated second-to-best results across all criteria; although, the ***SuperPrompt*** method did so for most of them. Also, somewhat unexpectedly, the ***Concepts*** method consistently underperformed, indicating the importance of LOs identification, as missing this step is the only difference between the ***Concepts*** method and the complete pipeline.

As the primary comparison against a conventional baseline, we tested a single planned contrast per criterion between ***LOs*** and ***Content***. For each criterion, the contrast was computed from the fitted LMM as the difference in estimated marginal means, $\Delta = \hat{\mu}_{LOs} - \hat{\mu}_{Content}$, where a positive Δ indicates higher expected ratings for the proposed pipeline. We report Δ with its 95% confidence interval and a two-sided Wald test p-value. To quantify practical magnitude on a standardized scale, we also report Hedges' g.

[4]A table summarizing criterion-specific flags for abnormal rating patterns is available in the online repository

Table 2
Marginal means per method (standard deviations in parentheses).

Criterion	*Content*	*Concepts*	*LOs*	*SuperPrompt*
Relevance	4.16 (0.89)	4.10 (0.93)	**4.44** (0.75)	4.14 (0.81)
Answerability	3.91 (1.31)	4.01 (1.23)	**4.63** (0.61)	4.14 (1.14)
Interestingness	3.47 (1.26)	3.12 (1.30)	**3.72** (1.12)	3.48 (1.15)
Importance	3.86 (1.03)	3.81 (1.00)	**4.21** (0.89)	3.88 (0.92)
Utility	3.68 (1.08)	3.57 (1.12)	**4.02** (1.06)	3.83 (1.04)

Table 3 shows that ***LOs*** significantly outperformed ***Content*** on *Answerability*, *Relevance*, and *Importance*. The differences in *Interestingness* and *Utility* are positive, but not significant.

Table 3
Contrast ***LOs*** and the baseline ***Content*** for each criterion.

Criterion	Diff	95% CI	p	g
Relevance	+0.28	[0.00, 0.56]	**.047**	+0.56
Answerability	+0.72	[0.24, 1.20]	**.003**	+1.16
Interestingness	+0.25	[-0.17, 0.56]	.248	+0.35
Importance	+0.35	[0.01, 0.75]	**.047**	+0.61
Utility	+0.34	[-0.14, 0.59]	.079	+0.53

To compare the four methods on an overall quality score, we combined the five quality criteria (Relevance, Answerability, Interestingness, Importance, and Utility). We first calculated an overall quality score for each expert's evaluation by averaging the available criterion scores. If a rating was filtered out during the data pre-processing, we averaged the remaining ones. Next, for each generated question, we averaged these overall quality scores across experts to obtain one score per question. For each method, we then took the mean of its questions' scores as the method's overall quality. Finally, we used a one-way ANOVA to test whether the mean overall quality differs across methods, and we followed up with three targeted comparisons between the experimental method and each alternative, using Holm correction across the three tests.

The one-way ANOVA on the overall quality score showed a significant difference between the methods, $F(3, 76) = 3.73$, $p = .014$, $\eta_p^2 = .13$. ***LOs*** had the highest mean for the overall quality ($M = 4.25$, 95% CI [4.06, 4.44]), followed by ***SuperPrompt*** ($M = 3.92$, 95% CI [3.71, 4.13]), ***Content*** ($M = 3.86$, 95% CI [3.61, 4.11]), and ***Concepts*** ($M = 3.80$, 95% CI [3.57, 4.02]), with $N = 20$ items per method. ***LOs*** scored significantly higher than all alternative settings in targeted post-hoc comparisons, including ***Content*** ($\Delta = 0.39$, $z = 2.63$, $p_{\text{Holm}} = .017$), ***Concepts*** ($\Delta = 0.45$, $z = 3.05$, $p_{\text{Holm}} = .007$) and ***SuperPrompt*** ($\Delta = 0.33$, $z = 2.21$, $p_{\text{Holm}} = .027$).

5.3. Difficulty and Cognitive Level

Experts evaluated how well the generated questions matched the *Difficulty* and *Cognitive Level* of the source textbooks sections using non-ordinal scales, therefore, these two factors have been evaluated only qualitatively.

Fig. 2(a) shows the *Difficulty* distribution per method. The difficulty of questions generated by the proposed strategy was rated as *Just right* (50%) of the times, which is the largest share among all conditions. It also had the smallest share of *Too easy* ratings - only (8%). Overall, in 92% of cases, the difficulty of questions generated by the experimental strategy was rated as either just right or within an small distance from it. For comparison, for ***SuperPrompt***-generated questions, this number is 88%, for ***Content*** - 84% and for ***Concepts*** - only 70%. In general, while we do not draw statistical conclusions, this distribution suggests that the proposed pedagogically-aligned prompt-chaining strategy is more likely to generate questions with appropriate difficulty.

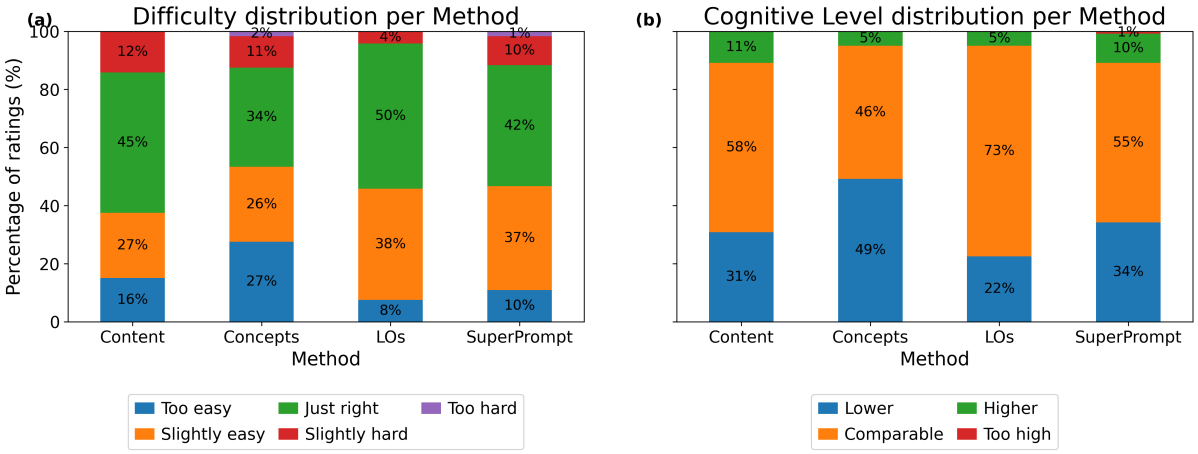

Figure 2: Distributions of (a) Difficulty ratings and (b) Cognitive Level labels by method (percentages over all collected ratings).

Fig. 2(b) summarizes the distribution fo experts' ratings regarding how well the *Cognitive Level* of generated questions matched the Cognitive Level of the material presented in the original section. Most ratings across conditions fell into *Comparable* and *Lower* categories. Again, the experimental strategy is a clear winner: 73% of the ratings indicated ana appropriate *Cognitive Level*, which strongly outperformed all other conditions. We want to underline the distinction between mismatching the *Difficulty* and the *Cognitive Level*. A question targeting a *Lower Cognitive Level* is qualitatively different from one that is merely *Slightly Easy*. While slightly easier questions may still support learning, questions operating at a lower cognitive level tend to trivialize the learning material and offer limited instructional value. For example, instead of asking learners to apply or reason about a concept, such questions may simply ask them to complete or recall a definition, thereby failing to engage the intended cognitive processes. Therefore a much lower percentage of such questions generated by the proposed strategy is another argument in favor of its effectiveness.

6. Conclusion

We examined whether organizing LLM-based question generation as a chain of pedagogically aligned prompts improves the quality of generated assessment items compared to prevalent monolithic prompting approaches. By explicitly decomposing the generation process into concept extraction, learning objective identification, and question construction, the proposed pipeline externalizes pedagogical reasoning that is otherwise left implicit in single-prompt methods.

Results from an expert-based evaluation indicate that the proposed strategy consistently produces higher-quality questions across multiple dimensions. While statistically significant improvements were most pronounced for Answerability, Relevance and Importance the pedagogically aligned pipeline achieved the highest overall quality scores and more reliably matched the intended difficulty and cognitive demands of the source material. In particular, explicitly modeling learning objectives emerged as a critical factor: the approach that omitted this step underperformed even when key concepts were provided. These findings suggest that consistent pedagogical alignment, rather than prompt complexity alone, is central to effective LLM-based question generation in educational contexts.

7. Limitations and Future Work

This research has several important limitations. First, the evaluation was based exclusively on expert judgments rather than learning outcomes; future work should examine how generated questions might affect student learning, engagement, and knowledge retention. Second, the experiments focused on

a single textbook in statistics, which challenges the generalizability of the obtained results. Further investigation should explore the efficacy of the approach in other domains, including those that are less structured than statistics. Third, although prompt chaining improves control and transparency, the results of intermediate steps have not been thoroughly analyzed and may still contain errors. The next steps in this direction should methodically investigate the quality of the extracted and weighted concepts as well as the quality of the generated LOs. Additionally, the pipeline's prompting strategy could be refined via automated prompt optimization, integrating reinforcement learning and validation-based evaluation loops. Finally, more types of assessment formats beyond multiple-choice and fill-in-the-blank can be added to the arsenal of generated questions.

Declaration on Generative AI

The authors have not employed Generative AI tools while writing this paper.

References

[1] W. Xing, N. Nixon, S. Crossley, P. Denny, A. Lan, J. Stamper, Z. Yu, The use of large language models in education, IJAIED 35 (2025) 439–443.

[2] H. McNichols, M. Zhang, A. Lan, Algebra error classification with large language models, in: AIED Conference, Springer, 2023, pp. 365–376.

[3] S. P. Neshaei, R. L. Davis, A. Hazimeh, B. Lazarevski, P. Dillenbourg, T. Käser, Towards modeling learner performance with large language models, in: EDM Conference, 2024, pp. 759–768.

[4] S. Sarsa, P. Denny, A. Hellas, J. Leinonen, Automatic generation of programming exercises and code explanations using large language models, in: Conference on international computing education research, 2022, pp. 27–43.

[5] I. Estévez-Ayres, P. Callejo, M. Á. Hombrados-Herrera, C. Alario-Hoyos, C. Delgado Kloos, Evaluation of llm tools for feedback generation in a course on concurrent programming, IJAIED 35 (2025) 774–790.

[6] E. Chen, J.-E. Lee, J. Lin, K. Koedinger, Gptutor: Great personalized tutor with large language models for personalized learning content generation, in: Learning@Scale Conference, 2024, pp. 539–541.

[7] D. Baradari, N. Kosmyna, O. Petrov, R. Kaplun, P. Maes, Neurochat: A neuroadaptive ai chatbot for customizing learning experiences, in: Conference on Conversational User Interfaces, ACM, New York, NY, USA, 2025.

[8] S. Sonkar, K. Ni, S. Chaudhary, R. Baraniuk, Pedagogical alignment of large language models, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 13641–13650.

[9] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al., Siren's song in the ai ocean: A survey on hallucination in large language models, Computational Linguistics 51 (2025) 1373–1418.

[10] I. Arcuschin, J. Janiak, R. Krzyzanowski, S. Rajamanoharan, N. Nanda, A. Conmy, Chain-of-thought reasoning in the wild is not always faithful, arXiv preprint arXiv:2503.08679 (2025).

[11] T. Wu, M. Terry, C. J. Cai, Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts, in: CHI conference on human factors in computing systems, 2022, pp. 1–22.

[12] H. W. Awalurahman, R. F. Aji, I. Budi, Transformer and large language models for automatic multiple-choice question generation: A systematic literature review, IEEE Access (2025).

[13] N. Mulla, P. Gharpure, Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications, Progress in Artificial Intelligence 12 (2023) 1–32.

[14] G. Biancini, A. Ferrato, C. Limongelli, Multiple-choice question generation using large language models: Methodology and educator insights, in: Conference on User Modeling, Adaptation and Personalization, 2024, pp. 584–590.

[15] U. Lee, H. Jung, Y. Jeon, Y. Sohn, W. Hwang, J. Moon, H. Kim, Few-shot is enough: exploring chatgpt prompt engineering method for automatic question generation in english education, Education and Information Technologies 29 (2024) 11483–11515.
[16] W. Chan, A. An, H. Davoudi, A case study on chatgpt question generation, in: International Conference on Big Data, IEEE, 2023, pp. 1647–1656.
[17] S. Maity, A. Deroy, S. Sarkar, Can large language models meet the challenge of generating school-level questions?, Computers and Education: AI 8 (2025) 100370.
[18] Q. Wang, R. Rose, N. Orita, A. Sugawara, Automated generation of multiple-choice cloze questions for assessing english vocabulary using gpt-turbo 3.5, in: Conference on NLP for Digital Humanities, 2023, pp. 52–61.
[19] K. W. Chan, F. Ali, J. Park, K. S. B. Sham, E. Y. T. Tan, F. W. C. Chong, K. Qian, G. K. Sze, Automatic item generation in various stem subjects using large language model prompting, Computers and Education: AI 8 (2025) 100344.
[20] H. A. Nguyen, S. Bhat, S. Moore, N. Bier, J. Stamper, Towards generalized methods for automatic question generation in educational domains, in: ECTEL Conference, Springer, 2022, pp. 272–284.
[21] D. Nuzzo, E. Vakaj, H. Saadany, E. Grishti, N. Mihindukulasooriya, Automated generation of competency questions using large language models and knowledge graphs, 3rd NLP4KGc@ SEMANTICs (2024).
[22] S. Al Faraby, A. Adiwijaya, A. Romadhony, Review on neural question generation for education purposes, IJAIED 34 (2024) 1008–1045.
[23] I. Alpizar-Chacon, S. Sosnovsky, Knowledge models from pdf textbooks, New Review of Hypermedia and Multimedia 27 (2021) 128–176.
[24] A. Kotha, J. Lee, E. Zakariasson, et al., Gpt-5 prompting guide, OpenAI Cookbook, 2025. URL: https://cookbook.openai.com/examples/gpt-5/gpt-5_prompting_guide.
[25] OpenAI, Prompt optimizer | openai api, OpenAI Platform Documentation, 2025. URL: https://platform.openai.com/docs/guides/prompt-optimizer.
[26] H. F. Rahmlow, K. K. Woodley, Objectives-based testing: A guide to effective test development, Educational Technology, 1979.
[27] F. Alonso, G. López, D. Manrique, J. M. Viñes, Learning objects, learning objectives and learning design, Innovations in education and teaching international 45 (2008) 389–400.
[28] D. R. Krathwohl, A revision of bloom's taxonomy: An overview, Theory into practice 41 (2002) 212–218.
[29] J. S. Nevid, N. McClelland, Using action verbs as learning outcomes: Applying bloom's taxonomy in measuring instructional objectives in introductory psychology., Journal of Education and Training Studies 1 (2013) 19–24.
[30] B. Illowsky, S. Dean, Introductory Statistics, OpenStax, Houston, TX, 2013. URL: https://openstax.org/books/introductory-statistics/pages/1-introduction.
[31] C. Van der Lee, A. Gatt, E. Van Miltenburg, E. Krahmer, Human evaluation of automatically generated text: Current trends and best practice guidelines, Computer Speech & Language 67 (2021) 101151.
[32] S. Elkins, E. Kochmar, I. Serban, J. C. Cheung, How useful are educational questions generated by large language models?, in: AIED Conference, Springer, 2023, pp. 536–542.
[33] S. Nakagawa, H. Schielzeth, Repeatability for gaussian and non-gaussian data: a practical guide for biologists, Biological Reviews 85 (2010) 935–956.