Toward Realistic AI-Generated Student Questions to Support Instructor Training

Francesco Cardia*,[1], Viviana Pentangelo*,[2][0009−0003−1425−9398], Stefano Lambiase[2][0000−0002−9933−6203], Carmine Gravino[2][0000−0002−4394−9035], Fabio Palomba[2][0000−0001−9337−5116], and Mirko Marras**,[1][0000−0003−1989−6057]

[1] University of Cagliari, Cagliari, Italy
f.cardia32@studenti.unica.it, mirko.marras@acm.org
[2] University of Salerno, Salerno, Italy
{vpentangelo, slambiase, gravino, fpalomba}@unisa.it

Abstract. Instructor effectiveness is fundamental to student learning, with the ability to manage student inquiries serving as a critical component of effective teaching. Student questions represent a valuable training resource for instructors to strengthen their teaching strategies, yet interactions with students are often constrained by several factors. In this paper, we investigate how instructors perceive machine- and student-generated questions, considering the potential for the former to complement the latter in a cost-effective manner. Our study involved 121 undergraduate students and an equivalent number of simulated students modeled using a state-of-the-art large language model, generating over 360 questions in total based on video lectures given by seven university instructors. We assessed whether instructors could distinguish between human- and machine-generated questions and how they evaluated their relevance, clarity, answerability, challenge level, and cognitive depth. Results show that instructors struggle to differentiate between the two sets of questions, with accuracy close to random chance. Instructors tended to (i) rate machine-generated questions slightly higher in relevance, clarity, answerability, and challenge—though only relevance and answerability showed significant differences—and (ii) associate them marginally more often with higher-order cognitive skills. This confirms the potential of machine-generated questions as tools for instructor training. **Repository**: https://github.com/tail-unica/realistic-ai-generated-questions.

Keywords: Generative AI · Instructor Training · Student Simulation · Student Questioning · Question Generation · Large Language Models.

1 Introduction

Motivation. Instructors can influence student learning and achievement, with some instructors showing greater effectiveness in fostering positive educational

* These authors contributed equally to this work.
** Corresponding author.

outcomes for students than others [1]. Teaching effectiveness encompasses actions that enhance or facilitate learning [26]. Improving this effectiveness requires identifying strategies that support instructors in optimizing their practices [17]. In this regard, a key aspect is answering students' questions and doubts during a lesson, which may represent a challenging situation for the instructors, demanding them to refine and reflect on their communication and content delivery.

In recent years, this challenge has been amplified by the spreading adoption of remote learning environments—such as Massive Open Online Courses (MOOCs) [7]. On the one hand, students are adopting this modality for its flexibility and availability, which are increasing access to educational resources, overcoming constraints such as class size, limited time, and student availability [11, 24]. On the other hand, they limit the instructor's ability to interact with students in real-time: student questions often arrive asynchronously, and instructors must find effective ways to respond post-hoc without immediate feedback or classroom cues [23]. As this model becomes increasingly common, supporting instructors in anticipating and handling student inquiries becomes essential.

Open Issues. In traditional settings, instructors may progressively improve their responses to student questions through direct interaction. However, in asynchronous online contexts, opportunities for such practice are scarce. In this way, many instructors are not regularly exposed to dynamic, question-driven teaching environments, which limits their ability to innovate or adjust methods based on immediate feedback [12, 31]. Critically, instructors have no way to predict what questions will be asked nor rehearse answering them, which could reduce their potential to improve student learning outcomes [32, 30]. From a constructivist view, this lack of questioning inhibits the reflective practice essential for instructor learning. Questions serve not only as indicators of student thinking but also as stimuli for instructors to construct pedagogical responses [25, 10].

To address these limitations, simulations have increasingly been introduced in educational contexts [15]. However, these efforts are generally directed at the instructor side of the educational interaction. They simulate how an instructor might respond to a student's input, such as through teaching assistants who offer feedback [19]. Rarely, though, do these simulations replicate students [3]. Most studies using learner simulations focus on method-related goals, e.g., benchmarking [27, 22] and validating techniques [29]. Some focus on optimizing student models via parameter learning [8]. However, no study has examined simulated student-generated questions nor instructor perceptions of them.

Contributions. In this paper, we want to take a foundational step toward using machine-generated student questions as a support mechanism for instructor training. The goal is to investigate whether it is feasible to simulate realistic student inquiries in view of their possible future usage for instructor training and continuous teaching improvement — particularly in asynchronous contexts such as MOOCs[3]. To this end, we recruited seven university instructors who recorded

[3] It should be noted that machine-generated questions should not be seen as a replacement for real student input, but as a complementary resource in settings where direct student interaction is sparse or delayed as well as for pre-class training.

video lectures on topics they regularly teach at the university. These lectures were then presented to real students, who formulated their questions, and a large language model generated inquiries based on student personas. Instructors then reviewed both sets of questions, attempting to distinguish between machine- and human-generated questions ($\mathbf{RQ}_1$), while also assessing their clarity, relevance, answerability, challenge level, and cognitive depth ($\mathbf{RQ}_2$).

The rest of this paper is organized as follows: Section 2 details the study design. Section 3 presents the findings, while Section 4 interprets them. Finally, Section 5 summarizes key insights and points to future research.

2 Methodology

In this section, we present the methodology designed to investigate the potential of machine-generated questions as a source that can complement real student inquiries in instructional settings where additional support is needed (Figure 1). Our methodological choices are informed by constructivist learning studies, which emphasize the importance of engaging instructors in interactions with learner-generated content [25, 10]. In this context, student questions — whether real or simulated — act as cognitive artifacts that challenge instructors to analyze, interpret, and respond in ways that foster their growth. With this in mind, we followed a three-step approach: (1) preparing teaching materials by collecting and curating lectures from university instructors; (2) generating student questions through two parallel processes—real students submitting inquiries and a large language model producing machine-generated questions using few-shot prompting with student personas; and (3) assessing instructor perceptions by evaluating their ability to distinguish between human- and machine-generated questions ($\mathbf{RQ}_1$) and rating questions based on key quality dimensions ($\mathbf{RQ}_2$).

2.1 Lectures Collection

When designing the study, we considered various instructional formats, including face-to-face lectures, live interactive sessions, and pre-recorded video lessons. While traditional and live lectures offer spontaneous interactions, they introduce substantial variability, such as unscripted peer discussions and real-time instructor adaptations, that complicate systematic comparisons across participants. In contrast, short, structured video lectures ensure that both real and simulated students engage with consistent instructional content, thereby enabling controlled and replicable evaluation of question-generation phenomena [21]. Importantly, this design choice is not a simplification but a deliberate alignment with the realities of modern MOOC and online learning platforms, where asynchronous video content is the norm and one of the key challenges is the lack of real-time student feedback [28]. With this educational setting as a target, we could enable, for instance, the use of machine-generated questions by instructors to anticipate common points of confusion, refine instructional content, and prepare for potential student inquiries in asynchronous environments (e.g., forums).

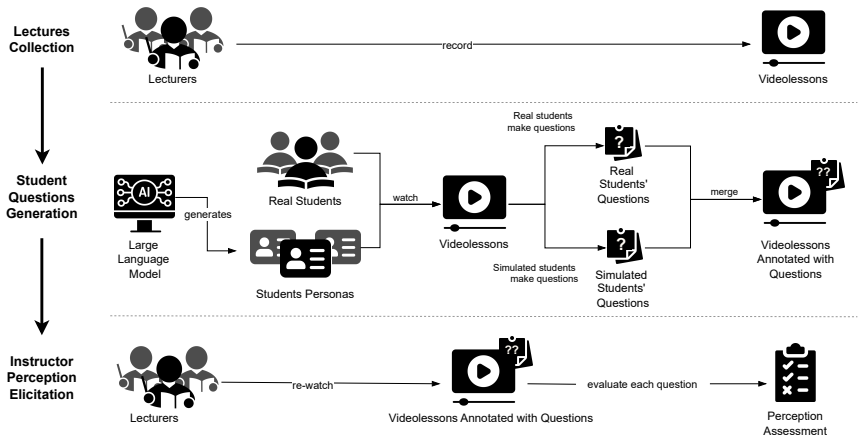

Fig. 1. Method. Overview of our method, including video lecture collection, question generation by real and simulated students, and instructor perceptions collection.

Table 1. Collected Lectures. Lectures, recorded by university instructors and covering distinct topics, served for generating human- and machine-generated questions.

ID	Topic	Complexity	Duration	Transcript	Frequency
SVM	Support Vector Machines	Beginner	9:45	1,469 words	151 words/min
NNE	Neural Networks	Beginner	11:14	1,588 words	141 words/min
FAI	Fairness in AI	Intermediate	12:01	1,718 words	143 words/min
IMC	Image Classification	Intermediate	12:18	1,796 words	146 words/min
TSC	Time Series Classification	Intermediate	8:36	1,087 words	126 words/min
TAR	Transformer Architecture	Advanced	8:17	1,300 words	157 words/min
AIC	AI and Code Smells	Advanced	9:55	1,195 words	120 words/min

In a first stage, we therefore curated a set of video lectures[4] that simulate real-world instructional scenarios typically part of online learning platforms. Specifically, we collected seven distinct lectures from as many instructors, each covering a unique topic relevant to a Computer Science Bachelor's degree program with a focus on Artificial Intelligence[5]. The instructors were not provided with a structured template or predefined slides, allowing them to maintain their own teaching styles and delivery. The cohort of instructors consisted of two females (28.6%) and five males (71.4%), ensuring a degree of perspectives diversity.

The collected lectures exhibit notable variability in duration, transcript length, and complexity (Table 1). The shortest lecture lasted just over 8 minutes (TAR, 8:17 min), while the longest exceeded 12 minutes (IMC, 12:18 min), reflecting dif-

[4] Once recorded, the lectures were uploaded to YouTube under private access settings, ensuring that they could only be accessed via direct links, for the subsequent steps.

[5] This choice enabled the analysis of question-generation patterns across diverse subtopics while minimizing confounding factors from disciplinary differences. Future work can extend this study to other domains to assess generalizability.

ferent depths of topic coverage. Transcript lengths also varied, from 1,087 words (`TSC`) to 1,796 words (`IMC`), showing differences in speaking pace and content density. The words per minute (WPM) further indicates a spectrum of instructional styles, with some instructors delivering fast-paced explanations (`TAR`, 157 WPM), while others adopt a slower approach (`AIC`, 120 WPM). In terms of topic complexity, the collected lectures include beginner (`SVM`, `NNE`), intermediate (`FAI`, `IMC`, `TSC`), and advanced (`TAR`, `AIC`) lectures. This diversity in content and delivery allows for a robust evaluation of real and simulated student questions.

2.2 Student Questions Generation

As a second step, we employed a dual, parallel approach: real students formulated their own questions while engaging with instructional content, while a large language model generated questions following structured student personas.

Human Questions Generation. The question-generation process was conducted asynchronously, allowing students to engage with the assigned videos at their own pace. This setting ensured that students interacted with the lectures in a flexible manner, similar to modern online course platforms. Participants were recruited through a voluntary survey, with a total of 121 students enrolled in the third year of the same Computer Science Bachelor's degree program from which the instructors and lecture recordings originated. This consistency in academic background facilitated a controlled analysis, ensuring that all participants had comparable foundational knowledge and were accustomed to the instructional style of their institution. The sample included 40% female students.

To ensure that student-generated questions were meaningful and reflective of genuine knowledge gaps, the recruitment survey asked participants to self-report their familiarity with each lecture topic. Based on these self-assessments, students were assigned to two lectures where they had the least prior knowledge. This assignment strategy balanced the distribution of participants across all lectures, ensuring that question-generation was not biased toward topics where students already had expertise and so possibly were less likely to make questions.

Following this assignment, students received an email with experimental instructions and access to a custom web-based platform we developed[6]. This platform facilitated structured question submission while ensuring precise tracking. Specifically, once submitted, each question was appended to a unique CSV file, recording (i) the date and time of the submission, (ii) the anonymized participant identifier, (iii) the lecture from which the question comes, (iv) the lecture time when the question was asked, and (v) the question text. For example, one entry might indicate that on June 12, 2024, at 11:35:16, participant 87 submitted a question for the lecture `TRA` at the 1:17 mark, asking: "Does this type of architecture form the basis of the most well-known translation applications?".

[6] First, students enter their anonymized ID, assigned video IDs, and perform a self-assessment on key student profiling dimensions. Subsequent pages display each video with an embedded player and an interface component for question submission.

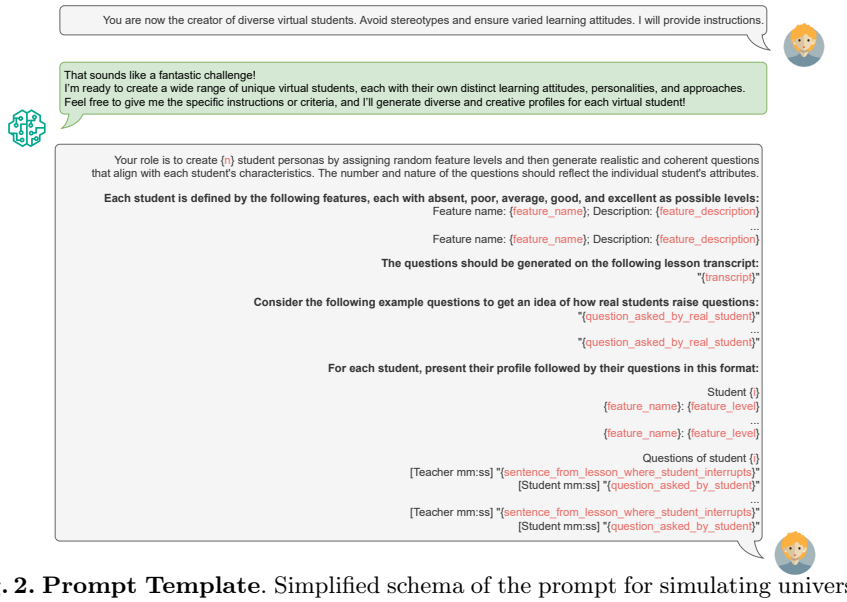

Fig. 2. Prompt Template. Simplified schema of the prompt for simulating university student personas and generating realistic questions based on randomized characteristics. These templates structure student features, their levels, and questioning moments.

In the instructions, students were encouraged to submit questions whenever they encountered confusion or curiosity, mimicking the natural inquiry process that occurs in real learning environments. To further promote active engagement and prevent passive participation, at the end of the session, they were also required to provide a short summary of each lecture they watched. These summaries, typically five to six sentences long, not only served as a means of verifying their comprehension but also reinforced content retention by encouraging students to synthesize and reflect on the key points presented in the lecture.

Machine Questions Generation. The first step involved extracting lecture transcripts, which were automatically generated by `YouTube`. These transcripts captured the content of each lecture from a machine perspective, differing from human students who had access to both audio and visual cues. This decision was made for three key reasons. First, it eliminates multimodal confounders such as intonation and gestures, ensuring that differences between machine- and human-generated questions stem from textual comprehension rather than extraneous factors. Second, it aligns with computational constraints, as most state-of-the-art models are trained primarily on text and lack robust multimodal reasoning. Third, prior research confirms that textual data alone captures essential semantic content for meaningful question generation [6]. While future work may explore multimodal approaches, arranging them goes beyond the scope of this study.

The generation process was grounded in two key pillars: (1) modern role-play prompting [18] to guide the language model's behavior, and (2) relevant student

profiling dimensions to create diverse and realistic student personas [9, 13, 2]. For each lecture, we employed a two-stage framework leveraging a role-playing prompting technique [18], adapted specifically to the educational domain and the student questioning task. The selection of this technique builds on prior findings showing that role-playing enhances large language models' zero-shot reasoning [18]. Specifically, we used GPT-4o as the language model, since the same prior work showed its superiority. The prompt we designed (see Figure 2) provided structured instructions, including persona's attributes, the lecture transcript, example questions from other contexts, and the expected output format.

Within the prompt, we asked to characterize each student persona by twelve dimensions grounded in findings from prior research in student modeling [2]. Such dimensions included prior knowledge, interest, motivation, concentration, critical thinking, ability to learn, and memory [9], as well as intuition, creativity, logical reasoning, emotionality, and language proficiency [13]. In GPT-4, simulated students were internally modeled by randomly assigning a level to each dimension on a discrete scale from 1 (absent) to 5 (excellent), with unit increments[7]. Such choice facilitated different questioning patterns. For instance, motivated and knowledgeable personas might produce in-depth inquiries, while less engaged personas might generate simpler or fewer questions.

The model instantiated each student persona and processed the lecture transcript in a sequential manner, identifying points at which questions would naturally emerge. The number of questions generated per persona remained unconstrained to allow for natural variation in questioning behavior. All generated questions were systematically recorded in a structured dataset formatted identically to the human-generated corpus, enabling direct comparability. Each entry was annotated with a timestamp reflecting the date and time of submission, an anonymized participant identifier, the corresponding lecture video, and the precise temporal location within the lecture when the question was raised. Additionally, the dataset preserved the textual content of each generated question.

2.3 Instructor Perception Elicitation

With the collected questions, we implemented a structured assessment protocol to determine whether instructors could reliably distinguish between human- and machine-generated questions and to analyze their perceptions regarding relevance, clarity, answerability, level of challenge, and cognitive depth.

To ensure an unbiased evaluation, instructors conducted their assessments[8] independently, without access to other raters' evaluations. Each instructor re-

[7] While these attributes exist on a continuum in real learners, categorizing them into discrete levels aligns with widely used educational frameworks, such as Bloom's taxonomy, which structure cognitive and affective traits into distinct stages.

[8] We acknowledge that evaluating questions in isolation does not fully reflect the complexity of classroom instruction, where instructors consider interpersonal cues, prior discussion threads, and the evolving classroom atmosphere. However, our goal at this stage was to establish whether machine-generated questions are perceived as realistic, i.e., a prerequisite before integrating them into high-variance environments.

ceived a spreadsheet containing a set of fields about questions of their own lecture[9], including the timestamp the question was posed in the lecture, the text of the question, and a source field where they were asked to indicate whether they believed the question was human- or machine-generated. The spreadsheet also incorporated four numerical evaluation criteria (relevance, clarity, answerability, and level of challenge for instructors) aligned with established frameworks on the quality of student-generated questions [31, 12]. Instructors rated each criterion on a 1 to 10 scale, with higher values indicating stronger alignment with the respective property. Each question was also classified in terms of perceived cognitive depth according to Bloom's Taxonomy by the instructor [5], selecting among remembering, understanding, applying, analyzing, evaluating, or creating.

To maintain a balanced assessment, we selected an equal number of human- and machine-generated questions to be added into the evaluation spreadsheet for a given lecture. The total number of questions per lecture was determined by the smaller of the two available sets (human- or machine-generated questions) for that specific lecture (i.e., down-sampling the larger set), to prevent imbalances and potential biases. The selected questions were randomly shuffled to eliminate any patterns that could indicate their origin in the spreadsheet. The inclusion of timestamps allowed instructors to evaluate each question in its proper lecture context. It should be noted that instructors were not informed that the machine-generated questions were produced solely from textual transcripts of the lectures.

3 Experimental Results

We analyzed the students' questions to determine whether instructors can reliably distinguish between human- and machine-generated inquiries ($\mathbf{RQ}_1$). Additionally, we examined how instructors perceive the questions in terms of relevance, clarity, answerability, challenge level, and cognitive depth ($\mathbf{RQ}_2$).

3.1 Turing Test for Question Source Classification [$\mathbf{RQ_1}$]

In this analysis, we were interested in evaluating whether instructors could reliably distinguish between machine-generated (MQ) and human-generated (HQ) questions. Specifically, we examined the extent to which instructors correctly classified MQ as machine-generated or misclassified them as human-generated, and similarly, whether HQ were correctly identified as human-generated or misclassified as machine-generated. Results are summarized in Table 2, which reports the percentage of (in)correctly classified questions across different lectures. The final row presents the averaged results across all lectures, providing a general measure of the instructors' ability in differentiating between the two sources.

[9] To strengthen the evidence and improve generalizability, future work will incorporate cross-instructor rating conditions and extend the study to instructors from other disciplines and with varying levels of experience with technology.

Table 2. [RQ$_1$] Question Source Classification. The proportion of machine- (MQ) and human-generated questions (HQ) that were either correctly classified or misclassified. The Accuracy column represents the sum of correctly classified MQ and HQ per lecture. Bold values indicate the highest percentage per column.

Lecture	# Questions	Instructor Accuracy	MQ predicted as		HQ predicted as	
			MQ	HQ	MQ	HQ
SVM	44× 2	40.9%	22.7%	27.3%	31.8%	18.2%
NNE	64× 2	45.3%	29.7%	20.3%	**34.4%**	15.6%
FAI	54× 2	42.6%	16.7%	33.3%	24.1%	25.9%
IMC	52× 2	**69.2%**	**32.7%**	17.3%	13.5%	**36.5%**
TSC	76× 2	39.5%	13.2%	**36.8%**	23.7%	26.3%
TAR	34× 2	61.8%	26.5%	23.5%	14.7%	35.3%
AIC	36× 2	50.0%	19.4%	30.6%	19.4%	30.6%
Average	52× 2	49.9%	23.0%	27.0%	23.1%	26.9%

The average classification accuracy of 49.9% suggests that instructors struggled to distinguish between machine- and human-generated questions, performing basically as a random guesser. Specifically, on average, instructors correctly identified 23.0% of MQ, while 27.0% of MQ were misclassified as HQ. Conversely, 23.1% of HQ were incorrectly labeled as MQ, while 26.9% were correctly identified as HQ. The fact that more MQ were misclassified as HQ than correctly identified implies that machine-generated questions often appeared natural enough to be perceived as human-authored. Likewise, the relatively high proportion of HQ misclassified as MQ indicates that certain student-generated questions may have exhibited features resembling algorithmically structured questions.

Accuracy varied significantly across different lectures, ranging from 39.5% (TSC) to 69.2% (IMC), revealing considerable disparities in instructors' ability to classify questions depending on the topic. In some cases, such as IMC and TSC, instructors exhibited higher accuracy in recognizing MQ, with 32.7% and 36.8% correctly classified as machine, respectively. Meanwhile, for RNN, the highest misclassification of HQ as MQ was recorded at 34.4%, suggesting that human-generated questions in this lecture may have shared stylistic or structural features with machine-generated ones. The highest overall accuracy in IMC (69.2%) suggests that the questions in this lecture may have had clearer distinguishing characteristics between human and machine origins. On the other hand, TSC (39.5%) and FAI (42.6%) had notably lower accuracy, indicating that MQ and HQ in these topics were more ambiguous, making classification more difficult. When examining patterns across lectures with similar characteristics (Table 1), we observed that beginner and intermediate topics consistently led to the lowest classification accuracy, often with reversed judgments, except for IMC. This may be due to the straightforward nature of simpler topics, leading both machine- and human-generated questions to converge toward generic, surface-level inquiries.

RQ$_1$. *Instructors substantially struggled to distinguish machine from human questions, particularly in lectures with lower instructional levels. This finding suggests machine-generated questions as credible real learner-like inquiries.*

Table 3. [RQ$_2$] Evaluation of Relevance, Clarity, Answerability, Challenge. The average ratings of machine- (MQ) and human-generated (HQ) questions across the four numerical dimensions. The final row reports the average score per column. Bold values indicate the highest score in each lecture for each dimension. Statistical significance analyses between MQ and HQ in each lecture for each dimension are based on Mann-Whitney U tests with Bonferroni correction (* $p < 0.01$, ** $p < 0.05$).

Lecture	Relevance		Clarity		Answerability		Challenge	
	MQ	HQ	MQ	HQ	MQ	HQ	MQ	HQ
SVM	**4.41**	3.59	**6.14**	4.95	**6.23**	5.27	3.18	**3.55**
NNE	**7.41**	7.12	**7.78**	7.47	**6.94**	6.69	**3.50**	3.22
FAI	**8.70***	7.37	**8.48***	7.04	**8.52***	7.11	**5.93**	4.89
IMC	6.96	**7.00**	**6.69**	6.65	**7.27**	7.04	**6.08**	5.73
TSC	**5.89**	5.79	6.32	**6.61**	**6.50**	6.13	5.08	**5.34**
TAR	**8.06****	5.88	**9.29**	8.24	**9.06**	8.29	**6.18***	2.82
AIC	**9.33**	8.50	**8.89**	8.67	**7.89**	6.83	5.44	**6.06**
Average	**7.11****	6.45	**7.47**	6.98	**7.34***	6.67	**4.97**	4.56

3.2 Multi-Dimensional Question Quality Perception [RQ$_2$]

In a second analysis, we examined how instructors perceived machine- (MQ) and human-generated (HQ) questions across relevance, clarity, answerability, challenge level, and cognitive level. To address this, we analyzed the average ratings[10] . of instructors for the first four dimensions (Table 3) as well as the cognitive classification of questions according to Bloom's taxonomy (Table 4).

On average, MQs received higher scores than HQs across all four evaluated numerical dimensions (relevance, clarity, answerability, and challenge level). Statistically significant differences emerged in relevance ($p < 0.05$) and answerability ($p < 0.01$), with MQs rated as more relevant (7.11 vs. 6.45) and easier to answer (7.34 vs. 6.67). The higher relevance may emerge from LLMs generating questions instantly, anchoring them closely to the target part of the lecture. In contrast, students often take longer to process information and submit questions later — unless pausing the video — leading to prompts less tied to the specific part. The gain in answerability likely reflects the LLMs' intrinsic knowledge base, enabling them to generate questions with a sense of the expected answer, unlike students who naturally ask questions without knowing the answer. While MQs also scored higher in clarity (7.47 vs. 6.98) and challenge (4.97 vs. 4.56), these differences were not statistically significant, indicating less consistent patterns.

Examining individual lectures reveals notable similarities between MQs and HQs. In most cases, ratings were closely matched — though MQs often scored slightly higher — indicating a comparable perceived quality. Significant differences were rare exceptions. In `FAI`, MQs significantly outperformed HQs in relevance ($p < 0.01$), clarity ($p < 0.01$), and answerability ($p < 0.01$). This could be attributed to the longer transcript compared to other lectures, providing the LLM with more context to generate well-aligned and coherent questions. Similarly, in `TAR`, MQs showed significant gains in relevance ($p < 0.05$) and challenge

[10] We also explored the nested structure of our data - questions within lectures, lectures within instructors - by fitting mixed-effects models that included random intercepts for instructors and lectures. We also computed rank-biserial correlations to estimate effect sizes. These analyses yielded patterns consistent with the primary results.

Table 4. [RQ$_2$] Evaluation of Cognitive Levels. The percentage of machine- (MQ) and human-generated (HQ) questions classified into each Bloom's category, namely *Rem* for *Remembering*, *Und* for *Understanding*, *App* for *Applying*, *Anl* for *Analyzing*, *Evl* for *Evaluating*, and *Crt* for *Creating*. The last row shows the average scores across all lectures. Bold values indicate the highest score in each lecture for MQ and HQ.

Lecture	MQ						HQ					
	Rem	Und	App	Anl	Evl	Crt	Rem	Und	App	Anl	Evl	Crt
SVM	**36.36%**	9.09%	4.55%	31.82%	18.18%	0.00%	**40.91%**	22.73%	9.09%	9.09%	9.09%	9.09%
NNE	6.25%	**34.38%**	21.88%	15.62%	21.88%	0.00%	15.62%	**31.25%**	28.12%	18.75%	6.25%	0.00%
FAI	14.81%	**29.63%**	11.11%	7.41%	14.81%	22.22%	18.52%	**25.93%**	7.41%	7.41%	25.93%	14.81%
IMC	**42.31%**	11.54%	7.69%	15.38%	7.69%	15.38%	**23.08%**	**23.08%**	0.00%	19.23%	19.23%	11.54%
TSC	**31.58%**	5.26%	26.32%	7.89%	18.42%	10.53%	**57.89%**	2.63%	10.53%	7.89%	5.26%	15.79%
TAR	**23.53%**	11.76%	**23.53%**	**23.53%**	11.76%	5.88%	**58.82%**	29.41%	5.88%	5.88%	0.00%	0.00%
AIC	0.00%	**55.56%**	11.11%	11.11%	16.67%	5.56%	0.00%	**61.11%**	0.00%	0.00%	38.89%	0.00%
Average	22.12%	**22.46%**	15.17%	16.11%	15.63%	8.51%	**30.69%**	28.02%	8.72%	9.75%	14.95%	7.32%

($p < 0.01$). The higher WPM rate might have contributed, as the denser information flow could make it harder for students to formulate immediate, well-targeted questions. In contrast, the LLM's ability to process the entire transcript at once likely enabled it to generate more relevant and complex questions.

In terms of cognitive skill classification, MQs were more frequently associated with higher-order cognitive levels compared to HQs. MQs were often categorized under Analyzing (16.11%) and Evaluating (15.63%), while HQs were predominantly assigned to lower-order categories such as Remembering (30.69%) and Understanding (28.02%). This suggests that MQs typically required deeper engagement and critical thinking. However, this pattern varied across lectures. For example, in `NNE`, both MQs and HQs were mostly categorized under Understanding (34.38% and 31.25%, respectively), indicating similar cognitive demands. In contrast, `SVM` showed a clear distinction, with 31.82% (9.09%) of MQs (HQs) classified as Analyzing, reinforcing the cognitive strength of MQs in this lecture.

RQ$_2$. *Instructors rated machine-generated questions higher than human-generated ones in relevance, clarity, answerability, and challenge level, but with statistically significant gains only in relevance and answerability. The former were also more frequently associated with higher-order cognitive skills.*

4 Discussion and Implications

Our findings reveal how instructors distinguish between machine-generated (MQ) and human-generated (HQ) questions and perceive their quality. This section explores the implications of and contextualize these results within prior works.

One notable outcome of this study is that instructors struggled to reliably distinguish between MQs and HQs (**RQ**$_1$). Misclassification rates were balanced in both directions, indicating that MQs were frequently mistaken for authentic student contributions, while HQs were often confused for machine-generated content. This result aligns with prior studies on the effectiveness of machine-generated content in educational settings, which have shown that well-designed

machine-generated questions can mimic human inquiry patterns and pass as plausible student-generated contributions [31, 29, 14]. For instructor training, this finding has significant implications. Since instructors exhibited no systematic accuracy in classifying MQs versus HQs, their perception of student engagement may not be significantly altered by the presence of machine-generated questions. This suggests that MQs could be incorporated into professional development programs without disrupting instructors' natural assessment of student inquiry. By simulation, MQs can help instructors practice adaptive teaching. For example, the metaverse can enhance training with simulated student avatars raising dynamic questions, exposing instructors to diverse questioning behaviors and refining responses to spontaneous inquiries [20, 4]. Importantly, we view machine-generated questions not merely as content artifacts but as pedagogical prompts designed to support instructor reflection. This aligns with constructivist views of teaching as a form of learning, where engaging with student inquiries helps instructors refine their mental models of effective instruction [25, 10].

Our findings also highlight how instructors perceive machine-generated questions (MQs) in terms of relevance, clarity, answerability, and challenge ($\mathbf{RQ}_2$). While MQs generally received higher ratings than HQs across these dimensions, statistically significant differences were observed only in select cases. This suggests that although MQs are often perceived as well-structured, content-aligned, and easy to answer based on the available instructional materials, these advantages are not consistently significant. Nonetheless, these findings align with prior work showing that machine-generated questions can show strong structural coherence and relevance [16, 24]. Our results also indicate that MQs were more frequently associated with higher-order thinking skills, such as analyzing and evaluating, aligning with previous studies suggesting that machine-generated questions tend to emphasize structured reasoning and analytical depth [30, 32]. This is a promising opportunity: since MQs are perceived as of comparable quality, they could serve as scaffolding tools to support instructor training.

While real student questions undoubtedly carry unique qualities, such as informal phrasing or misconceptions, that reflect authentic cognitive processes, our goal was to examine the extent to which MQs can approximate the overall quality of student questioning. Rather than replacing authentic interactions, MQs should be viewed as valuable complements that enhance instructor preparedness, especially in contexts where direct student input may be limited, delayed, or inconsistent, such as in asynchronous or large-scale online learning environments. Far from narrowing pedagogical awareness, exposure to a spectrum of MQs (including also higher-order and reflective prompts) can enrich instructors' repertoire of responses and promote anticipatory thinking. At the same time, incorporating a balance of question types by playing with the simulated student personas in MQ design can help maintain sensitivity to the full range of student learning needs. More importantly, MQs offer a low-risk, high-availability resource for rehearsal and reflection, empowering instructors to refine their pedagogical strategies and better support students across varying levels of understanding.

5 Conclusion and Future Work

In this paper, we examined instructors' ability to distinguish between machine- and human-generated questions and their perceptions of question quality. To this end, we introduced a novel protocol for comparing real and machine-generated student inquiries, a novel method for generating student-like questions using structured personas, and an novel annotated dataset including human- and machine-generated questions to video lectures we collected as well. Results show that instructors struggled to differentiate between the two, indicating that artificially generated questions closely resemble authentic student inquiries. Moreover, machine-generated questions were consistently rated higher in relevance, clarity, and answerability and were more frequently classified under higher-order cognitive skills. These findings show their potential in instructor training to refine questioning strategies, improve assessment design, and foster critical thinking.

While our findings show the potential of machine-generated questions in instructor training, several limitations suggest avenues for future research. The sample of instructors and curricula leaves space for investigating the generalizability of our results. However, our study serves as a foundational step in exploring the role of machine-generated questions in instructional settings. We prioritized internal validity by selecting a coherent curricular context. This allowed us to control for domain-specific variables and establish a clear methodological framework. In future work, we will extend this analysis to other disciplines — especially those with distinct characteristics from computer science — but we view our contribution as a necessary step toward broader interdisciplinary validation. In future work, we also plan to systematically compare question outputs across diverse personas, e.g., varying in prior knowledge, confidence, or motivation, and analyze how these factors shape the style, complexity, and pedagogical relevance of generated questions. Furthermore, while the machine-generated questions were based solely on textual transcripts, students engaged with the full audiovisual content of the lectures. This intentional design choice allowed us to isolate and assess the capabilities of text-based models in simulating student-like questioning. In future work we will explore multimodal approaches that incorporate visual and auditory information to more closely mirror the full context available to students. Our current findings thus provide a necessary baseline for evaluating the added value of such multimodal integration. To enhance realism, future studies will also consider sampling simulated student personas from survey data rather than uniform random draws, better aligning with authentic student diversity.

Beyond dataset representativeness, instructor evaluations may have been influenced by covariates, such as the tendency to rate more polished questions higher regardless of their pedagogical depth. Implementing explicit controls for such covariates and collecting instructors beliefs on machine-generated questions could provide a more comprehensive assessment. Moreover, our findings show that machine-generated questions can closely resemble those posed by students, that is an important step in validating their potential role in educational settings. In future iterations, we aim to move beyond isolated evaluation by embedding machine-generated student questions within real instruction contexts,

e.g., with controlled A/B tests that compare instructor responses, reflection level, and improvements across conditions involving no questions, student questions, and machine-generated questions. Our current design thus represents a foundational, low-risk step to validate realism and instructional plausibility before advancing to this deployment in the wild. Finally, evaluating multiple large language models would offer a deeper understanding of how different generative architectures, training data, and decoding strategies influence the formulation and quality of machine-generated questions. Such a comparative analysis could reveal important nuances in how models represent student thinking, generate various cognitive levels of inquiry, or reflect domain-specific patterns. This technical direction will be also actively pursued in future steps to enhance robustness and generalizability of machine-generated questions across educational settings.

References

1. Bardach, L., Klassen, R.M.: Smart teachers, successful students? a systematic review of the literature on teachers' cognitive abilities and teacher effectiveness. Educational Research Review **30**, 100312 (2020)
2. Batool, S., Rashid, J., Nisar, M.W., Kim, J., Kwon, H.Y., Hussain, A.: Educational data mining to predict students' academic performance: A survey study. Education and Information Technologies **28**(1), 905–971 (2023)
3. Bhowmik, S., West, L., Barrett, A., Zhang, N., Dai, C.P., Sokolikj, Z., Southerland, S., Yuan, X., Ke, F.: Evaluation of an llm-powered student agent for teacher training. In: Proc. of ECTEL 2024. pp. 68–74. Springer (2024)
4. Bilotti, U., Di Dario, D., Palomba, F., Gravino, C., Sibilio, M.: Machine learning for educational metaverse: How far are we? In: Proc. of ICCE 2023. pp. 01–02. IEEE (2023)
5. Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., Krathwohl, D.R., et al.: Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain. Longman New York (1956)
6. Blšták, M., Rozinajová, V.: Automatic question generation based on sentence structure analysis using machine learning approach. Natural Language Engineering **28**(4), 487–517 (2022)
7. Calvo, S., Lyon, F., Morales, A., Wade, J.: Educating at scale for sustainable development and social enterprise growth: The impact of online learning and a massive open online course (mooc). Sustainability **12**(8), 3247 (2020)
8. Chen, H., Wang, J.: Awaking the slides: A tuning-free and knowledge-regulated ai tutoring system via language model coordination. arXiv preprint arXiv:2409.07372 (2024)
9. Chrysafiadi, K., Virvou, M.: Student modeling approaches: A literature review for the last decade. Expert Systems with Applications **40**(11), 4715–4729 (2013)
10. Clarke, D., Hollingsworth, H.: Elaborating a model of teacher professional growth. Teaching and teacher education **18**(8), 947–967 (2002)
11. Graesser, A.C., Person, N.K., Magliano, J.P.: Question asking during tutoring. American Educational Research Journal **31**(1), 104–137 (1994)
12. Griffith, J., Vercellotti, M.L., Folkers, H.: What's in a question? a comparison of student questions in two learning spaces. Teaching and Learning in Communication Sciences & Disorders **3**(1), 7 (2019)

13. Hamim, T., Benabbou, F., Sael, N.: Student profile modeling: an overview model. In: Proc. of ICSCA 2019. pp. 1–9 (2019)
14. Huang, Y., Wang, L.: A case study on chatgpt question generation. IEEE Transactions on Learning Technologies (2023)
15. Käser, T., Alexandron, G.: Simulated learners in educational technology: A systematic literature review and a turing-like test. International Journal of Artificial Intelligence in Education **34**(2), 545–585 (2024)
16. Kaya, S., Temiz, M.: Improving the quality of student questions in primary science classrooms. Journal of Baltic Science Education **17**(5), 800–811 (2018)
17. Klassen, R.M., Kim, L.E., Rushby, J.V., Bardach, L.: Can we improve how we screen applicants for initial teacher education? Teaching and Teacher Education **87**, 102949 (2020)
18. Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., Dong, X.: Better zero-shot reasoning with role-play prompting. In: Proc. of NAACL 2024. pp. 4099–4113 (2024)
19. Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., Käser, T.: Ai or human? evaluating student feedback perceptions in higher education. In: Proc. of ECTEL 2024. pp. 284–298. Springer (2024)
20. Pentangelo, V., Di Dario, D., Lambiase, S., Ferrucci, F., Gravino, C., Palomba, F.: Senem: A software engineering-enabled educational metaverse. Information and Software Technology p. 107512 (2024)
21. Poehner, M.E.: Dynamic assessment in the classroom. The Concise Companion to Language Assessment p. 55 (2024)
22. Shao, Y., Li, L., Dai, J., Qiu, X.: Character-llm: A trainable agent for role-playing. In: Proc. of EMNLP 2023. pp. 13153–13187 (2023)
23. Sher, A.: Assessing the relationship of student-instructor and student-student interaction to student learning and satisfaction in web-based online learning environment. Journal of Interactive Online Learning **8**(2) (2009)
24. Song, D.: Student-generated questioning and quality questions: A literature review. Research Journal of Educational Studies and Review **2**(5), 58–70 (2016)
25. Van Es, E.A., Sherin, M.G.: Learning to notice: Scaffolding new teachers' interpretations of classroom interactions. JTTE **10**(4), 571–596 (2002)
26. Vieluf, S., Klieme, E.: Teaching effectiveness revisited through the lens of practice theories. In: Theorizing teaching: Current status and open issues, pp. 57–95. Springer International Publishing Cham (2023)
27. Wang, N., Peng, Z., Que, H., Liu, J., Zhou, W., Wu, Y., Guo, H., Gan, R., Ni, Z., Yang, J., et al.: Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In: Findings of the Association for Computational Linguistics ACL 2024. pp. 14743–14777 (2024)
28. Wang, R., Cao, J., Xu, Y., Li, Y.: Learning engagement in massive open online courses: A systematic review. In: Frontiers in Education. vol. 7, p. 1074435 (2022)
29. Wu, L., Kim, S.: Generative students: Using llm-simulated student profiles to support question item evaluation. Proc. of the ACM L@S 2024 (2024)
30. Xiang, W., Liu, X., Wu, Y., Chen, Y.: Khanq: A dataset for generating deep questions in education. In: Proc. of COLING 2022 (2022)
31. Xu, Y., Zhao, M., Liu, X., Sun, X., Zhang, H.: Can autograding of student-generated questions quality by chatgpt match human experts? IEEE Transactions on Learning Technologies (2024)
32. Yang, L., Chen, X., Liu, B., Gao, J.: Learningq: A large-scale dataset for educational question generation. In: Proc. of ICWSM 2018 (2018)