



Instructors' perceptions of a large language model framework for generating automated multiple-choice questions from video transcripts

Giacomo Cassano¹ · Daniel Ebbert^{2,3} · Negin Mirriahi² · Nicoletta Di Blas¹

Received: 10 June 2025 / Accepted: 31 May 2026
© The Author(s) 2026

Abstract

Video-Based Learning has become a central modality in digital education, with embedded Multiple-Choice Questions (MCQs) recognized as an effective strategy for promoting engagement and assessing comprehension. However, manually creating high-quality MCQs is challenging for instructors. This study presents a framework that leverages Large Language Models (LLM) to automate MCQ generation from video transcripts while ensuring alignment with established item-writing criteria. Unlike most prior approaches, the framework explicitly integrates the Item-Writing Flaws (IWF) checklist into the generation prompt, rather than using it solely for post-hoc evaluation. It segments video transcripts into topics, extracts and classifies key concepts, and applies structured prompting strategies to generate MCQs adhering to the IWF checklist and additional criteria from the literature, including relevance, difficulty, contextual specificity, and answerability. The framework was evaluated through a mixed-methods study involving seven instructors from diverse disciplines who assessed 257 generated MCQs. The results indicate that the majority of items met established quality guidelines. Specifically, 84.05% of the questions were rated as clear, 86.38% contained plausible distractors, and 93.39% were directly answerable based on the video content. The main issues identified concerned relevance, reflecting occasional difficulty in distinguishing illustrative examples from core instructional content, and linguistic precision, with a tendency to overgeneralize instructors' statements. In semi-structured interviews, instructors reported that the framework could substantially reduce MCQ preparation time while requiring only minor refinements to improve question specificity. Overall, the findings suggest that embedding pedagogical quality principles directly into LLM prompting represents a viable and promising direction for assessment automation.

Keywords Video-based learning · Multiple-choice questions · Large-language models · Artificial Intelligence in education · Assessment automation · Teacher perception

Introduction

The use of Multiple-Choice Questions (MCQs) in Video-Based Learning (VBL) is emerging as an effective pedagogical practice (Brame, 2016; Kimbrel & Gantner, 2021). They can be embedded within videos or used as formative assessments, enhancing engagement while providing immediate insights into learner comprehension (Kay & Knaack, 2009; Mayer, 2017).

Creating high-quality MCQs is a labor-intensive process that involves aligning questions with learning objectives and adhering to established guidelines for quality, as defined in the literature (Haladyna & Downing, 1989; Haladyna et al., 2002). High-quality MCQs require content accuracy and the elimination of common flaws, such as ambiguous wording, poorly constructed distractors, or unintentional clues that might compromise the validity of the assessment (Haladyna et al., 2002). Research highlights that many instructors face challenges in dedicating sufficient time and resources to this task, due to competing responsibilities, such as course preparation, grading, and administrative duties (Haladyna et al., 2002). To address these challenges, studies are emerging leveraging artificial intelligence (AI) to automate the generation of MCQs (e.g. Arif et al., 2024; Bhowmick et al., 2023). While promising, current AI-generated MCQs often fall short of the quality required for effective pedagogy: studies have highlighted some limitations, such as poor contextual relevance, and difficulties in generating plausible distractors (Arif et al., 2024; Grévisse, 2024). These shortcomings suggest that while AI can assist in reducing the time burden on instructors, it cannot yet replace them in the creation of MCQs.

This study presents a framework that leverages Large Language Models (LLMs) to automate MCQ generation from video transcripts while ensuring alignment with established item-writing criteria. Unlike most prior approaches, which apply quality guidelines primarily for post-hoc evaluation, the framework explicitly integrates the Item-Writing Flaws (IWF) checklist directly into the generation prompt. Its contribution lies in operationalizing these pedagogical quality principles within the LLM-based generation pipeline itself. It adopts a mixed-methods evaluative design: after developing the framework, its effectiveness is examined through instructors' assessments of the generated MCQs, combining quantitative survey data with qualitative insights from semi-structured interviews. Accordingly, this study is guided by the following research question:

How do instructors perceive the quality of multiple-choice questions generated by an LLM-based framework that explicitly integrates item-writing quality principles?

By focusing on instructors' perceptions, this question allows us to examine the extent to which the generated MCQs align with educators' pedagogical expectations and instructional standards.

State of the art

Multiple-choice questions and their quality

MCQs, introduced in the early twentieth century alongside standardized testing in the U.S., have become a widely used assessment tool due to their efficiency and scalability. They enable rapid evaluation across a broad range of topics, support objective grading, and are

well-suited for automated scoring, making them particularly effective for large-scale educational assessments (Haladyna, 2004; Macy et al., 2010).

A well-designed MCQ test must meet established quality criteria to eliminate Construct-Irrelevant Variance, i.e. extraneous factors that distort the accurate assessment of the intended skill or knowledge (Haladyna & Downing, 2004). For this purpose, Haladyna and Downing (1989) developed a taxonomy of guidelines for writing MCQs that included 31 guidelines. In their revision of these guidelines they organized them into five categories: content concerns, formatting concerns, style concerns, writing the stem, and item wording (Haladyna et al., 2002). In recent years, these guidelines have been revisited and frequently used to validate MCQs. Among their adaptations, the Item-Writing Flaws (IWF) rubric (Moore et al., 2023) has gained particular prominence; it comprises 19 standardized criteria (see Appendix A). This rubric has been validated through several studies (Breakall et al., 2019; Moore et al., 2021; Rush et al., 2016), and the IWF checklist is now the most commonly used framework for the validation of MCQs in research studies (Arif et al., 2024; Moore et al., 2023; Nedeau-Cayo et al., 2013).

Automatic evaluation of MCQ quality before LLMs

Thanks to advancements in machine learning algorithms, it has become possible to develop fully automatic methods for evaluating the quality of MCQs. The initial methods for automatic MCQ evaluation were based on Natural Language Processing (NLP) algorithms and metrics like BLEU and METEOR.

BLEU (Bilingual Evaluation Understudy) is a metric used to evaluate the quality of text generated by machine learning models, particularly in NLP tasks such as machine translation. It measures the similarity between machine-generated text and one or more reference texts written by humans (Papineni et al., 2002). The BLEU score has occasionally been used in the evaluation of MCQs to assess the linguistic similarity between automatically generated questions and reference questions written by experts.

Metric for Evaluation of Translation with Explicit ORdering (METEOR) is another automatic evaluation metric used primarily to measure the similarity between machine-generated text and human reference text. METEOR is more suitable than BLEU for evaluating the linguistic quality of automatically generated MCQs, thanks to its ability to recognize synonyms and syntactic variations (Banerjee & Lavie, 2005).

However, using methods like BLEU and METEOR is not sufficient to assess the overall quality of MCQs because they do not consider the pedagogical content of the question but focus solely on the lexical similarity between human-created and machine-generated content. Moreover, these metrics often do not correlate well with human evaluations (van der Lee et al., 2021).

Some approaches assess MCQ quality using classification models trained on large datasets of student responses, analyzing indices such as item difficulty and discrimination (Haladyna, 2004). However, these metrics are calculated post-hoc, relying on test performance data rather than informing the design of items in advance. Other studies employ rule-based methods to detect specific formal flaws, such as implausible distractors or structural inconsistencies (van der Waa et al., 2021; Wang et al., 2021). Although these methods provide interpretable and transparent signals, they primarily target surface-level characteristics or retrospective indicators of quality. Overall, while such approaches have advanced automated MCQ analysis, they do not directly embed

established pedagogical quality principles into the item generation process itself. This limitation highlights a gap that has more recently been addressed by LLM-based approaches.

Generation of MCQs with LLMs

The literature contains a few studies that report attempts to use LLMs for generating MCQs from educational materials. Elkins et al. (2023) conducted a human evaluation of educational questions generated using LLMs with few-shot prompting, engaging teachers to assess their quality based on Relevance, Grammatical Correctness, Answerability, and Adherence. Bhowmick et al. (2023) proposed a modular framework for generating MCQs from textual content, including separate components for question generation, answer prediction, and distractor generation. Their evaluation combined automated metrics with human judgments assessing question quality, answer correctness, groundedness in the source content, and distractor plausibility. Similarly, Cheung et al. (2023) focused on generating MCQs from medical textbooks, employing expert evaluators to assess questions based on Appropriateness, Clarity, Specificity, Relevance, Discriminatory Power of Alternatives, and Suitability for Medical Examinations. In the mathematical domain, Li et al. (2024) and Hang et al. (2024) developed MCQ generation models tailored to contextual mathematical reasoning. They assessed the questions according to Contextual Appropriateness, Rationality of the Options, Rationality of the Stem, and Generative Potential. Hang et al. (2024) presented MCQ-Gen, a generative AI framework for the automated generation of MCQs based on GPT-4o (OpenAI, 2024a). They evaluated the generated MCQs with the following set of criteria: Grammatical Fluidity, Answerability, Diversity, Complexity, and Relevance.

The study most closely related to our work is that of Arif et al. (2024). Their methodology segments video transcripts into key moments, following Bhat et al. (2023), and generates questions based on transcript summaries and titles provided as contextual input to the LLM. For quality assessment, they rely on the 19-item Item-Writing Flaws rubric (Moore et al., 2021), supplemented with additional context-based criteria tailored to video-based learning. The generated MCQs are assessed both by human raters and by LLMs (GPT-3.5 and GPT-4). Overall, their findings indicate generally positive results, particularly with respect to relevance, answerability, and contextual specificity.

As can be seen, research on LLM-based MCQ generation is still in its early stages. The existing studies are heterogeneous, differing in prompting strategies as well as in the criteria used to assess the quality of the generated questions. Despite this variability, findings generally indicate promising levels of quality (Bhowmick et al., 2023; Elkins et al., 2023; Li et al., 2024), and a consistent benefit reported across studies is the reduction in time required for instructors to produce assessment items (Bhowmick et al., 2023; Cheung et al., 2023). At the same time, limitations remain. Arif et al. (2024) identify persistent challenges in generating plausible distractors, which constitute the most frequent flaw in their analysis, along with less frequent issues related to contextual grounding.

Notably, while several studies employ selected Item-Writing Flaws (IWF) criteria for post-hoc evaluation, none explicitly integrate these established quality principles into the generation prompt itself. This gap is addressed in the present study.

The framework

We developed a LLM-based framework that aims to generate a comprehensive set of MCQs from an educational video, ensuring that all relevant concepts covered in the video are represented. The process is divided into four sequential steps: (1) the video is processed to generate a transcript, (2) the transcript is segmented into topics, (3) each topic is analyzed to extract a set of concepts categorized into classes, and (4) MCQs are generated from these concepts. Figure 1 shows a graphical breakdown of the framework.

The framework leverages GPT-4o (OpenAI, 2024a) for concept identification and MCQ generation. The model was not trained or fine-tuned by the authors; instead, reliability is addressed at the level of use by guiding the LLM through structured prompts and explicit task constraints. LLM interactions are managed through LangChain (LangChain, 2024), which structures inputs, parameters, and expected outputs for each invocation. The model returns responses in JSON format, ensuring consistency and facilitating downstream processing.

Transcript extraction and pre-processing

The first step performed by the system is the extraction of the transcript from the video. This is achieved using Whisper, an advanced speech-to-text model developed by OpenAI (OpenAI, 2024b). Specifically, the system employs the Whisper-1 (v2-large) model to extract text with both word-level and segment-level precision. Word precision ensures that the exact timing of each word is captured with millisecond accuracy, while segment precision is essential for maintaining the punctuation and structure of the spoken content, which is not provided by word-level extraction alone.

The system then aggregates words into segments based on punctuation marks (e.g., periods, commas, ellipses, question marks, exclamation marks). This process ensures that each segment corresponds to a complete sentence and that no sentence is split across multiple segments. The output of this step is a transcript consisting of a set of time-ordered, non-overlapping segments, each containing a full sentence and bound by a start and end time, measured in milliseconds.

Topics extraction

The second step takes the processed transcript as input and generates a time-ordered list of adjacent topics (each topic starts immediately after the previous one). This marks the first invocation of the LLM, managed through a chain in LangChain. This step is necessary to

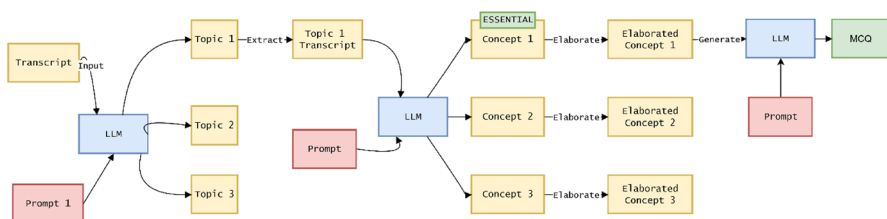


Fig. 1 Graphical breakdown of the framework

break the work into smaller tasks, especially considering that the transcript of a video can be very long and, therefore, cannot be processed using a single prompt. The chain organizes the interaction with the LLM, supplying a prompt, input parameters, and formatting instructions to ensure consistent output. The input includes the transcript formatted with time intervals for each segment (e.g., “[start second]—[end second]: text”).

The LLM is prompted to analyze the transcript, identify distinct topics, and define the start and end times for each based on the time intervals of the segments. Every sentence is assigned to a specific topic, ensuring there are no overlaps. The output is a structured list where each topic is characterized by a name summarizing its content, start and end timestamps in HH:MM:SS format, and a brief description derived from the corresponding transcript segments.

Concepts extraction and classification

The third step involves concept extraction, where the transcript for each topic is processed to categorize individual concepts. The further segmentation of topics into concepts is essential to enable classification that allows the system to focus on key concepts and formulate relevant questions. This step also employs the LLM through a chain in LangChain, using a prompt designed to guide the model in extracting concepts and assigning them to predefined categories. The transcript input is formatted as time-bound segments, where each row specifies the start and end time along with the corresponding text (e.g., “start second-end second: text”).

The prompt instructs the LLM to segment the transcript into semantically coherent instructional units, defined as complete ideas conveyed by one or more related sentences (e.g., a definition, explanation, or example). The segmentation follows the natural structure of the video and is intentionally designed to ensure comprehensive coverage of the instructor’s explicitly stated content, rather than to model fine-grained pedagogical distinctions or conceptual relationships. Moreover, the framework does not explicitly model common student misconceptions or predefined pedagogical taxonomies, as it relies exclusively on the instructor’s articulated content.

Each concept is assigned a title, the corresponding transcript excerpt, a rewritten standalone explanation, timestamps, and a functional category. Concepts are classified as *Essential* (core ideas central to the instructional objectives), *Supporting* (examples or clarifications), or *Organizational* (structural elements such as transitions). This categorization is deliberately coarse-grained and serves to separate core content from auxiliary or structural elements. The output is a structured list of concepts linked to their respective topics, forming the basis for subsequent MCQ generation.

MCQ generation

In the final step, the LLM generates one MCQ for each essential concept, guided by a prompt explicitly grounded in the IWF checklist. Each item includes a self-contained stem, one correct answer, and three plausible distractors. The prompt operationalizes the IWF criteria by embedding them directly into the generation instructions, constraining the model to avoid common flaws such as ambiguous stems, implausible distractors, multiple correct answers, unintended cues, and negative or overly complex wording.

Each option is accompanied by concise LLM-generated formative feedback. For the correct answer, the feedback briefly reinforces the underlying concept by restating the key

rationale. For incorrect options, it clarifies why the choice is inappropriate, such as by indicating conceptual inaccuracies or misalignment with the video content, without introducing new information beyond what is needed to explain the error. Overall, the framework ensures that the generated MCQs are pedagogically sound and effectively cover the lesson's essential concepts.

MCQ evaluation method

Participants and context

We recruited seven faculty members from Politecnico di Milano (Italy), representing different disciplines, to assess the suitability of the MCQs generated by the framework. In total, 257 MCQs derived from 45 MOOC videos were reviewed. Instructors provided informed consent for the use of their recorded video lessons for research purposes.

The evaluation was conducted as an expert judgment procedure: each MCQ was assessed by a single faculty member, specifically the instructor who had authored and delivered the corresponding video. This design ensured that each item was evaluated by the original Subject Matter Expert responsible for the instructional content. Given the study's focus on disciplinary alignment, the course instructor was considered best positioned to determine whether the generated questions accurately reflected the intended concepts and learning objectives.

It must be noted that the videos are relatively short (averaging 7 min each) and are carefully scripted; furthermore, unlike traditional classroom recordings, they are free from interruptions and background noise, resulting in clean and accurate transcripts. Testing the framework with these high-quality videos represents an ideal case, as it allows for evaluation under optimal conditions. By minimizing external variables, this approach isolates framework-specific issues and identifies potential problems before testing in more complex environments like standard lecture recordings.

The instructors and their videos spanned diverse disciplines (see Table 1). This disciplinary diversity broadens the range of instructional contexts examined and provides preliminary evidence that the framework can be applied across different academic fields, including humanities, architecture, and engineering. Table 1 summarizes the videos and corresponding MCQs evaluated by each instructor. Across the 45 videos analyzed, which averaged 7.01 min in length, the framework generated a mean of 5.77 MCQs per video ($SD=2.65$).

Table 1 Overview of videos and generated questions by instructor and topic

Instructor	Topic	#Videos	#MCQ
1	Communication skills	10	68
2	Introduction to AI	10	46
3	Architecture	6	23
4	Machine Learning	4	39
5	Project Management	6	41
6	Design	4	25
7	Math	5	15

MCQ evaluation through questionnaire

The quantitative assessment of MCQ quality is based on the 19-item IWF checklist (see Appendix), supplemented by four content-related criteria derived from prior research (Arif et al., 2024; Bhowmick et al., 2023; Elkins et al., 2023): Relevance, Difficulty, Contextual Specificity, and Answerability.

Of the 19 IWF criteria, 16 concern formal or structural aspects of item construction (e.g., avoiding making the correct answer the longest option) and can be verified without subject-matter expertise. Four of these (criteria 3, 5, 14, and 16) are proactively prevented by the framework during generation (e.g., true/false formats are not allowed) and therefore do not require manual checking. The remaining 12 formal criteria were reviewed by the first and second authors through item-by-item inspection. In contrast, the instructors evaluated the three IWF criteria requiring expert judgment (4 – Ambiguous or unclear information; 6 – Implausible distractors; 18 – More than one or no correct answer), together with four content-related criteria (Relevance, Difficulty, Contextual Specificity, and Answerability). We also introduced an additional criterion, *Correctness*, applied when a question contained substantive flaws not captured by the preceding categories and deemed unacceptable by the instructor. These criteria formed the basis of the survey used to guide instructor evaluations (see Table 2).

Evaluation sessions were conducted online between each instructor and the researchers, during which the assessment criteria were explained and the LLM-generated MCQs were presented for review. Instructors rated each item according to the specified criteria and could provide optional comments, which were transcribed for qualitative analysis. Flaws were recorded in an Excel worksheet, with one row per question and one column per criterion. We analyzed the frequency and co-occurrence of flaws, as well as the distribution of questions by number of flaws (none, one, two, or three or more).

MCQ evaluation through Semi-Structured Interviews

To complement the quantitative assessment, semi-structured interviews were conducted with all seven instructors to gather qualitative feedback on the generated MCQs and their potential instructional use. The interview protocol included open-ended questions addressing (1) formal quality, (2) difficulty, (3) content coverage, (4) teaching utility (including willingness to use the items), and (5) overall impressions. Interviews were recorded, transcribed, and analyzed using thematic analysis. One researcher coded the transcripts according to the five predefined categories, identifying key themes to structure the analysis of strengths, limitations, and perceived applicability of the framework.

MCQ evaluation results

This section presents the quantitative and qualitative findings concerning instructors' assessments of the generated MCQs.

Survey

Table 3 summarizes the instructors' ratings of the 257 MCQs. The analysis begins with the criterion of *Correctness*, which functioned as a preliminary filter: questions judged

Table 2 Evaluation Criteria Used in the Instructor Survey for Assessing LLM-Generated MCQs

Criterion	Question	Options
<i>More than one or no correct answer</i>	Does the question have one, and only one, correct answer?	Yes/No
<i>Clarity</i>	Is the question clear and unambiguous?	Clear/Not clear
<i>Distractor Plausibility</i>	Are the distractors plausible?	Yes/No
<i>Relevance</i>	To what extent does the question address a concept that is relevant to the video lesson?	Low/Medium/High
<i>Difficulty</i>	How challenging is the question?	Easy/Fair/Difficult
<i>Contextual Specificity</i>	Does the question rely on specific content not available to the learner, making it unanswerable without the missing context?	Yes (undesirable)/No (desirable)
<i>Answerability</i>	Does the source text directly support the answer to the question?	Direct/Indirect/None

Table 3 Descriptive statistics of MCQ evaluation

Criteria	Category	Count	Percentage (%)
Correctness	Totally Wrong	15	5.84
	Evaluated	242	94.17
IWF 4 (Clarity)	Clear	216	84.05
	Not Clear	26	10.12
IWF 6 (Distractor Plausibility)	Plausible	222	86.38
	Not Plausible	20	7.78
IWF 18 (More than one or no correct answer)	Verified	8	3.11
	Not verified	234	91.05
Relevance	High	133	51.75
	Medium	60	23.35
	Low	49	19.07
Difficulty	Difficult	18	7.00
	Fair	152	59.14
	Easy	72	28.02
Contextual Specificity	No (Desirable)	217	84.44
	Yes (Undesirable)	25	9.73
Answerability	Direct	240	93.39
	Indirect	2	0.78
IWF 1 (Longest option is correct)	Verified	1	0.39
	Not verified	241	93.77
IWF 13 (Unfocused stem)	Verified	80	31.13
	Not verified	162	63.04
Remaining IWF criteria	Verified	0	0
	Not verified	242	94.16

unsuitable for use were not further evaluated on the remaining criteria. As noted earlier, this category captured substantive flaws not encompassed by the other evaluation dimensions and is discussed in greater detail in the following section. Following this, the three IWF criteria evaluated by the instructors (Item 4 – Clarity, Item 6 – Distractors plausibility, Item 18 – More than one or no correct answer) are presented. Next are the four content-related criteria: relevance, difficulty, contextual specificity, and answerability. Finally, the evaluation concludes with the 16 formal IWF criteria, which were assessed by the researchers.

Correctness: Fifteen MCQs (5.84%) were judged incorrect by the instructors for reasons not attributable to any IWF criteria. Consequently, the detailed criteria analysis was conducted on the remaining 242 items (although percentages reported in the table refer to the full set, including these fifteen “incorrect” cases). To investigate these instances, we examined the corresponding generation steps and found that the issues originated from inaccuracies in the initial concept extraction. We present a notable example in Table 4: The LLM correctly identifies a contrast between two approaches (“descriptive and educational”) and a third one (“problematic”). However, although three approaches are mentioned, the key point lies in the two conceptual groupings. This distinction is lost during MCQ generation, where the three approaches are treated as independent and equivalent categories.

Table 4 Example of incorrect MCQ generation flow

Transcript	Concept	MCQ
The main question is whether we need to actively engage the reader's interest or not. Descriptive and educational approaches assume that the reader is already interested and that it is sufficient simply to describe the work. This is, for example, the approach typically adopted by students. In contrast, what we might call problematic approaches take on the responsibility of generating the reader's interest and demonstrating why the work is valuable	The discussion contrasts two approaches to engaging readers: descriptive and educational approaches, which assume inherent reader interest, and problematic approaches, which actively seek to engage the reader by demonstrating the value of the work	What are the two main approaches to engaging readers discussed? A) Descriptive and educational approaches B) Narrative and analytical approaches C) Persuasive and informative approaches D) Creative and critical approaches

Clarity (IWF 4): Among the 257 evaluated MCQs, 26 (10.12%) were marked as “Not Clear.” Teachers identified multiple reasons contributing to the lack of clarity. A recurring theme is poor precision, flagged in 7 instances, where the phrasing of the question was too generic or ambiguous. Comments highlighted missing phrases or missing specifications. In multiple cases, teachers flagged misinterpretation of examples, where aspects mentioned as supporting examples were incorrectly treated as the main focus of the question. Additionally, inappropriate terminology was identified three times due to the AI translating terms using synonyms, whereas the technical term should have been preserved. For example, the word ‘trajectory’ was used instead of the correct term ‘strategy.’ Repeated phrasing or redundant questions occurred in two instances, and transcription errors (e.g., “SPATIAL” confused with “SPECIAL”) were noted once.

Distractor Plausibility (IWF 6): Twenty MCQs (7.78%) were judged to contain implausible distractors, typically because options were overly obvious or irrelevant. As distractor quality appeared related to perceived difficulty, a contingency analysis was conducted (see Table 5). Results indicate that 80% of questions with implausible distractors were rated as Easy, whereas items with plausible distractors showed a more balanced distribution across difficulty levels. This pattern suggests that weak distractors reduce item challenge.

More than one or no correct answer (IWF 18): IWF 18 occurs when the framework generates only wrong answers or more than one correct answer. In 8 cases (3.08%), this criterion was not met.

Relevance: Of the 257 questions, 133 (51.75%) were rated as highly relevant and 60 (23.35%) as moderately relevant, while 49 (19.07%) showed low relevance. As relevance depends on the framework’s concept classification, the results indicate generally effective identification of key ideas. However, some false positives remain, producing technically accurate questions that target peripheral rather than central content. Instructors noted that certain items were based on illustrative examples instead of core concepts, as the framework sometimes focused on the literal details of examples rather than their intended instructional meaning, resulting in superficially relevant but misaligned MCQs.

Difficulty: Eighteen MCQs (7.00%) were rated as difficult, 152 (59.14%) as fair, and 72 (28.02%) as easy. Instructor feedback suggests that difficulty was influenced not only by distractor quality (discussed above), but also by question focus and phrasing. Items perceived as too difficult often relied on minor details or used complex wording, leading to ambiguity or reduced clarity.

Contextual Specificity: This criterion assessed whether each question could be answered based solely on the information provided in the video. Twenty-five items (9.73%) were judged to lack essential contextual details, whereas 217 (84.44%) were considered sufficiently self-contained. In most problematic cases, the issue originated in the rewriting step of concept generation, where key contextual references were occasionally omitted. For example, the question “*What does the statement suggest about the love of mothers for their children?*” appears grammatically complete but becomes unanswerable because the referenced “statement” is not specified.

Table 5 Contingency table of distractor plausibility vs. difficulty

Plausibility/Difficulty	1 (Easy)	2 (Fair)	3 (Difficult)
Not plausible	16	4	0
Plausible	56	148	18

Answerability: Most questions (93.39%) were directly answerable using only the video content, with just two cases (0.78%) indirectly answerable due to misinterpretation during concept extraction and rewriting. In these instances, key details were altered, preventing a complete answer based solely on the transcript. Overall, answerability does not appear to be a significant issue in the current version of the framework.

Remaining IWF criteria: Among the formal IWF criteria checked by the researchers, only two were violated: IWF 1 (Longest option is correct) and IWF 13 (Unfocused stem). This suggests that most formal aspects of item design were adequately handled by the framework.

IWF 1 occurred in a single case (0.39%), where the correct answer was substantially longer than the distractors. IWF 13 was observed in 31.13% of cases, most frequently due to the use of an indefinite article (e.g., “a key factor”) instead of a definite one, potentially implying multiple correct answers. Although formally flawed, instructors indicated that such questions would generally remain usable in practice.

Performance across multiple metrics

This section analyzes how different flaws co-occur within the 242 MCQs deemed correct. For each item, we calculated the total number of flaws by aggregating all unmet criteria recorded in the evaluation worksheet.

IWF criteria, together with contextual specificity and answerability, were treated as binary variables, where a negative rating indicated a flaw. In contrast, relevance and difficulty were assessed on three-point scales and were not inherently considered flawed at their extremes, as items of varying relevance or difficulty may be intentionally included in instructional design. Consequently, these two dimensions were excluded from the flaw-count analysis.

For each question, we counted violations across the IWF criteria and the two additional binary dimensions. An item was considered flawed whenever at least one of these criteria was unmet. Overall, 121 questions (47.08%) met all criteria, 208 (80.93%) failed at most one criterion, and 238 (92.60%) failed fewer than three criteria, with only four MCQs exhibiting three or more flaws.

We conclude with a final consideration. Among the 80 MCQs that violated IWF 13 (“unfocused stem”), instructors confirmed that these items would generally remain usable in a testing context despite the formal flaw. In most cases (64 out of 80), the issue stemmed from the use of the indefinite article “a” or similar expressions, which, as noted earlier, may imply the existence of multiple possible answers. The count of flaws for each MCQ was therefore repeated without considering IWF 13, obviously giving vent to more positive results as can be seen in Table 6.

Semi-structured interview

Insights from semi-structured interviews with instructors were analyzed thematically, revealing five key themes: quality, coverage relevance, difficulty, usability, and overall impressions.

Quality. The quality of the MCQs generated by the LLM-based framework was generally perceived as positive, with instructors highlighting its potential to accelerate exam production and facilitate the creation of plausible and contextually appropriate questions. Considering the use of MCQs as a tool for formative or even summative assessment, *after*

Table 6 Descriptive statistics about the co-occurrence of flaws

	Considering all flaws	Not considering IWF 13
Pass all metrics	121	174
Fail one or no	208	231
Fail less than 3	238	239
Fail 3 or more	4	3
Avg failure	0.66	0.34

video viewing, several instructors noted that the system could significantly reduce the workload associated with the MCQ creation, especially in situations where multiple versions of questions are required. The quality of distractors was widely praised, with several instructors emphasizing that the distractors were coherent with the context. The precision of the questions was also addressed, with instructors highlighting that while the system succeeded in generating coherent and contextually appropriate questions, it occasionally produced items that were too broad or imprecise. This was particularly observed in questions related to more open-ended topics. Despite these limitations, the instructors considered the questions adequate.

Coverage and Relevance. Overall, the framework showed good performance in identifying key concepts and covering essential topics. While some instructors noted occasional inclusion of peripheral elements or omissions, often linked to visually conveyed content, most agreed that the primary learning objectives were adequately represented.

Difficulty. Some instructors found the questions occasionally too easy, highlighting the need for future refinement, as difficulty was not explicitly controlled.

Usage. Instructors saw strong potential for the system in exam and quiz creation. While manual refinement is still needed, they noted significant time savings, up to 80%, compared to traditional methods, highlighting efficiency as a major advantage.

Final Impressions. The final reflections and suggestions for improvement provided by the instructors pointed to potential enhancements for the framework. Some instructors expressed interest in testing the system with more technical courses or having a similar system for open-ended questions, while others proposed evaluating its performance when generating questions from slide-based presentations. Additional suggestions included incorporating support for multiple correct answers, which could foster logical reasoning and the ability to discern correct options.

Discussion

Framework and overall performance

This study introduced a structured pipeline for generating MCQs from video transcripts by operationalizing established item-writing principles and embedding the IWF checklist directly into the generation prompt. In addressing our research question “*How do instructors perceive the quality of multiple-choice questions generated by an LLM-based framework that explicitly integrates item-writing quality principles?*”, the findings indicate that

most items met established quality standards, with only a small proportion exhibiting multiple flaws.

Instructors' evaluations were generally positive across key pedagogical dimensions, particularly clarity, distractor plausibility, and answerability. Qualitative feedback further suggested that the generated questions were largely usable in practice and perceived as potentially reducing quiz preparation time, as instructors could focus on reviewing and refining items rather than composing them from scratch.

At the same time, several framework-level challenges emerged. A recurring issue concerned stem focus (IWF 13), often linked to subtle linguistic choices that introduced ambiguity. In particular, article usage occasionally led to implicit multiplicity (“a”), affecting the exclusivity of the correct answer, or unintended definitiveness (“the”). In some cases, this resulted in questions framed in more absolute terms than the original instructional discourse, for example by transforming a nuanced explanation into a question such as “What is THE primary characteristic of X?”, thereby amplifying the definitiveness of the original statement. While many such items remained usable, this pattern highlights the sensitivity of single best-answer formats to minor phrasing variations and suggests a tendency of the LLM to over-specify instructor statements when operationalizing them as assessment items. Additional issues included occasional losses of contextual specificity, misclassification of illustrative examples as essential concepts, and structural repetitiveness in question formulation. These findings suggest that, although the framework performs reliably at the level of formal compliance, further refinement is needed to enhance precision.

The observed strengths and weaknesses of the generated MCQs also allow for a reflection on the types of knowledge that appear to be more or less amenable to LLM-based question generation. Items targeting declarative or explicitly stated conceptual knowledge generally resulted in clearer single best-answer structures. By contrast, questions intended to assess context-dependent, procedural, or inferential knowledge more often exhibited ambiguity or overlapping answer options. These patterns suggest that, while LLMs can effectively transform surface-level conceptual content into assessable items, they face challenges when the target knowledge depends on situated context or implicit reasoning. This consideration helps clarify the kinds of assessment tasks for which additional instructional scaffolding may be particularly important.

Comparative analysis with related work

While acknowledging differences in evaluation methodologies across studies, comparison remains useful for situating our findings within the broader literature. Among the reviewed works, Arif et al. (2024) is the most closely aligned with our approach, as it also generates MCQs directly from video content. Their study reports recurring challenges related to contextual grounding and distractor plausibility. Although differences in evaluation procedures limit direct comparison, our results suggest comparatively stronger performance in these dimensions.

Other studies typically generate questions from pre-selected textual materials rather than autonomously extracted instructional content. For example, Elkins et al. (2023) and Cheung et al. (2023) rely on curated textual sources, while Bhowmick et al. (2023) generate questions from extracted snippets or topic-level inputs. In contrast, our framework identifies relevant instructional concepts directly from video transcripts, introducing an additional content-selection layer prior to question generation.

Li et al. (2024) adopt a structured retrieval-based approach, integrating externally sourced knowledge into mathematical MCQs. By comparison, our framework focuses on extracting and operationalizing concepts from within the original instructional material itself.

Taken together, these distinctions clarify the specific methodological contribution of our study within the evolving landscape of LLM-based MCQ generation.

Limitations

The present study was intentionally designed to rely on expert judgment rather than learner response data. While such judgments are central to instructional alignment, this choice limits the analysis to instructor-based evaluations. Future research could complement this approach with quantitative psychometric analyses (e.g., item difficulty and discrimination indices) to strengthen external validity.

A second limitation concerns the evaluation procedure: each set of MCQs was reviewed by the instructor who authored the corresponding video. Although this ensured close alignment with instructional intent, it may have introduced a familiarity bias, as instructors' deep knowledge of their own material could have influenced their interpretation of ambiguous or context-dependent questions. Future research could incorporate cross-instructor or blind evaluations to mitigate this bias and provide a more independent assessment of item quality.

Finally, the analysis relied exclusively on transcript-based representations of the videos. Content conveyed primarily through visual or multimodal elements (e.g., diagrams or on-screen text) was therefore not captured. Consequently, the generated MCQs reflect only verbally expressed material, highlighting an important boundary of the current framework. Integrating multimodal inputs represents a relevant direction for future development.

Conclusions and future directions

This study presented a structured framework for generating MCQs from instructional video transcripts using LLMs and examined instructors' perceptions of their quality. To the best of our knowledge, it is the first to explicitly embed the IWF checklist directly into the generation prompt as a guiding mechanism rather than using it solely for post-hoc evaluation. The findings suggest that systematically operationalizing established item-writing principles within the item generation process can support the production of items that align with recognized quality standards in the majority of cases.

In relation to our research question, how instructors perceive the quality of MCQs generated by an LLM-based framework that integrates item-writing principles, the results indicate generally positive perceptions. Most items were judged to meet key quality criteria, and only a limited proportion exhibited multiple flaws. Instructors also reported that the generated questions were largely usable in practice and perceived as potentially reducing the time required for quiz preparation.

When situated within the existing literature, the proposed framework addresses several recurring challenges reported in prior studies by embedding quality principles directly into the generation stage. Although methodological differences preclude direct comparison, our results suggest comparatively strong performance in areas such as distractor plausibility

and contextual grounding. These comparisons, however, should be interpreted as indicative rather than conclusive.

At the same time, several directions for further development emerged. First, the framework occasionally struggled to distinguish illustrative examples from core conceptual content, leading to overly fact-based items tied to peripheral details. Second, the LLM sometimes framed questions in more absolute terms than the original instructional discourse, reinforcing definitiveness where nuance was intended. Third, the current transcript-based approach excludes visually conveyed information, limiting the system to verbally expressed content. Future work integrating multimodal inputs may expand the range of assessable learning objectives. Finally, evaluating the framework across different LLM architectures may provide further insight into how model characteristics influence question quality.

Overall, the findings indicate that embedding pedagogical quality principles directly into LLM prompting represents a viable and promising direction for assessment automation. Continued refinement of both concept extraction and linguistic control mechanisms will be essential to further enhance the contextual precision, epistemic alignment, and instructional usefulness of automatically generated MCQs.

Appendix

Evaluation criterion	Description
Relevance	Does the question address a relevant concept from the video lesson? (Three levels: Low, Medium, High)
Difficulty	Does the question present an appropriate level of challenge? (Three levels: Easy, Fair, Difficult)
Contextual specificity	Does the question require information not provided in the lesson, making it unanswerable without additional context? (Binary: Yes – undesirable, No – desirable)
Answerability	Does the question have its answer directly supported by the lesson content? (Three levels: Direct, Indirect, None)
IWF 1: Longest option is correct	Avoid making the correct option consistently the longest, as this may signal the answer. (Binary: Yes – undesirable, No – desirable)
IWF 2: Gratuitous information in stem	Avoid including unnecessary information in the question stem. (Binary: Yes – undesirable, No – desirable)
IWF 3: True/false question	The options should not be a series of true/false statements
IWF 4: Ambiguous or unclear information	Questions and all options should be written in clear, unambiguous language
IWF 5: Convergence cues	Avoid convergence cues in options where there are different combinations of multiple components to the answer
IWF 6: Implausible distractors	Make all distractors plausible as good items depend on having effective distractors
IWF 7: Use of none of the above	Avoid “none of the above” options as it only really measures students ability to detect incorrect answers

Evaluation criterion	Description
IWF 8: Logical cues in stem	Avoid clues in the stem and the correct option that can help the test-wise student to identify the correct option
IWF 9: Use of all of the above	Avoid “all of the above” options as students can guess correct responses based on partial information
IWF 10: Fill-in-blank	Avoid omitting words in the middle of the stem that students must insert from the options provided
IWF 11: Absolute terms (never, always)	Avoid the use of absolute terms (e.g. never, always, all) in the options as students are aware that they are almost always false
IWF 12: Word repeats in stem and correct answer	Avoid similarly worded stems and correct responses or words repeated in the stem and correct response
IWF 13: Unfocused stem	The stem should present a clear and focused question that can be understood and answered without looking at the options
IWF 14: Complex or K-type	Avoid questions that have a range of correct responses, that ask students to select from a number of possible combinations of the responses
IWF 15: Grammatical cues in stem	All options should be grammatically consistent with the stem and should be parallel in style and form
IWF 16: Lost sequence in presentation of data	All options should be arranged in chronological or numerical order
IWF 17: Vague terms (sometimes, frequently)	Avoid the use of vague terms (e.g. frequently, occasionally) in the options as there is seldom agreement on their actual meaning
IWF 18: More than one or no correct answer	In single best-answer form, questions should have 1, and only 1, best answer
IWF 19: Negative worded stem (not, incorrect, except)	Negatively worded stems are less likely to measure important learning outcomes and can confuse students

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

Data availability The datasets generated and/or analysed during the current study are not publicly available due to privacy and consent restrictions but are available from the corresponding author on reasonable request.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Arif, T., Asthana, S., & Collins-Thompson, K. (2024). Generation and assessment of multiple-choice questions from video transcripts using large language models. In *Proceedings of the eleventh ACM Conference on learning @ scale*, 530–534. ACM.
- Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, (pp. 65–72). <https://aclanthology.org/W05-0909.pdf>
- Bhat, M. M., Meng, R., Liu, Y., Zhou, Y., & Yavuz, S. (2023). *Investigating answerability of LLMs for long-form question answering*. arXiv preprint [arXiv:2309.08210](https://arxiv.org/abs/2309.08210). <https://arxiv.org/abs/2309.08210>
- Bhowmick, A. K., Jagmohan, A., Vempaty, A., Dey, P., Hall, L., Hartman, J., Kokku, R., & Maheshwari, H. (2023). Automating question generation from educational text. In M. Bramer & F. Stahl (Eds.), *Artificial Intelligence XL* (Vol. 14381, p. 437). Springer Nature Switzerland.
- Brame, C. J. (2016). Effective educational videos: Principles and guidelines for maximizing student learning from video content. *CBE—Life Sciences Education*, 15(4), e6. <https://doi.org/10.1187/cbe.16-03-0125>
- Breakall, J., Randles, C., & Tasker, R. (2019). Development and use of a multiple-choice item writing flaws evaluation instrument in the context of general chemistry. *Chemistry Education Research and Practice*, 20(2), 369–382. <https://doi.org/10.1039/C8RP00262B>
- Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., Wong, R., & Co, M. T.-H. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PLOS ONE*, 18(8), e0290691. <https://doi.org/10.1371/journal.pone.0290691>
- Elkins, S., Kochmar, E., Serban, I., & Cheung, J. C. K. (2023). How useful are educational questions generated by large language models? In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial Intelligence in Education: Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky* (Vol. 1831, pp. 536–542). Springer Nature Switzerland.
- Grévisse, C. (2024). Comparative quality analysis of GPT-based multiple choice question generation. In H. Florez & M. Leon (Eds.), *Applied Informatics* (pp. 435–447). Springer Nature Switzerland.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Routledge.
- Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 37–50. https://doi.org/10.1207/s15324818ame0201_3
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17–27. <https://doi.org/10.1111/j.1745-3992.2004.tb00149.x>
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5
- Hang, C. N., Wei Tan, C., & Yu, P.-D. (2024). MCQGen: A large language model-driven MCQ generator for personalized learning. *IEEE Access*, 12, 102261–102273. <https://doi.org/10.1109/ACCESS.2024.3420709>
- Kay, R., & Knaack, L. (2009). Exploring the use of audience response systems in secondary school science classrooms. *Journal of Science Education and Technology*, 18(5), 382–392. <https://doi.org/10.1007/s10956-009-9153-7>
- Kimbrel, L. A., & Gantner, M. W. (2021). Student perceptions of instructor made videos with quizzes in an asynchronous online course. *International Journal of Educational Leadership Preparation*, 16(1), 24–44.
- LangChain. (2024). *LangChain* [Computer software]. <https://www.langchain.com/>
- Li, R., Wang, Y., Zheng, C., Jiang, Y.-H., & Jiang, B. (2024). Generating contextualized mathematics multiple-choice questions utilizing large language models. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education: Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky* (pp. 494–501). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64315-6_48
- Macy, M., Thorndike-Christ, T., & Lin, Y.-C. (2010). Parental reports of perceived assessment utility: A comparison of authentic and conventional approaches. *Infants & Young Children*, 23(4), 286–296. <https://doi.org/10.1097/IYC.0b013e3181f1ec1b>
- Mayer, R. E. (2017). Using multimedia for e-learning. *Journal of Computer Assisted Learning*, 33(5), 403–423. <https://doi.org/10.1111/jcal.12197>

- Moore, S., Nguyen, H. A., Chen, T., & Stamper, J. (2023). Assessing the quality of multiple-choice questions using GPT-4 and rule-based methods. In O. Viberg, I. Jivet, P. J. Muñoz-Merino, M. Perifanou, & T. Papathoma (Eds.), *Responsive and sustainable educational futures* (Vol. 14200, pp. 229–245). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-42682-7_16
- Moore, S., Nguyen, H. A., & Stamper, J. (2021). Examining the effects of student participation and performance on the quality of learnersourcing multiple-choice questions. In *Proceedings of the eighth ACM conference on learning @ scale*, (pp. 209–220) <https://doi.org/10.1145/3430895.3460140>
- Nedeau-Cayo, R., Laughlin, D., Rus, L., & Hall, J. (2013). Assessment of item-writing flaws in multiple-choice questions. *Journal for Nurses in Professional Development*, 29(2), 52–57. <https://doi.org/10.1097/NND.0b013e318286c2f1>
- OpenAI. (2024a). *ChatGPT* (GPT-4o) [Large language model]. <https://openai.com/2025-release-notes/#gpt-4o>
- OpenAI. (2024b). *Whisper* [Computer software]. <https://openai.com/research/whisper/>
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318)
- Rush, B. R., Rankin, D. C., & White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education*, 16(1), 250. <https://doi.org/10.1186/s12909-016-0773-3>
- van der Lee, C., Gatt, A., van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67, Article 101151. <https://doi.org/10.1016/j.csl.2020.101151>
- van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291, Article 103404. <https://doi.org/10.1016/j.artint.2020.103404>
- Wang, Z., Zhang, W., Liu, N., & Wang, J. (2021). Scalable rule-based representation learning for interpretable classification. *Advances in Neural Information Processing Systems*, 34, 30479–30491. <https://doi.org/10.48550/arXiv.2109.15103>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Giacomo Cassano received his PhD in Information Technology from Politecnico di Milano in 2025. His research focused on video-based learning, student engagement, learning analytics, and AI-supported educational technologies. He is currently affiliated with HOC-LAB at Politecnico di Milano, where he contributes to teaching activities in educational technology.

Daniel Ebbert is a Research Fellow at Adelaide University, affiliated with the Centre for Change and Complexity in Learning and the School of Computer Science and Information Technology. His research explores mind wandering, self-regulated learning, and video-based learning, with a particular focus on learners' responses to attentional lapses. He is currently investigating educational technologies that support cognitive and metacognitive processes during learning.

Negin Mirriahi is an Associate Professor and Academic Lead (Digital Innovation) in the School of Education at Adelaide University, co-Director of the Higher Education Research Network and member of the Centre for Change and Complexity in Learning. Her research focuses on learning analytics, artificial intelligence in education, video-based learning, self-regulated learning, blended and online learning, and academic staff development.

Nicoletta Di Blas is Full Professor at Politecnico di Milano and Scientific Director of HOC-LAB. Her research investigates how digital technologies can support teaching and learning by making student responses visible, interpretable, and actionable. Her work focuses on video-based learning, mixed reality, and AI-supported educational environments, and is grounded in research conducted in large-scale, authentic learning settings.

Authors and Affiliations

Giacomo Cassano¹  · **Daniel Ebbert**^{2,3} · **Negin Mirriahi**² · **Nicoletta Di Blas**¹

✉ Nicoletta Di Blas
nicoletta.diblas@polimi.it

Giacomo Cassano
giacomo.cassano@polimi.it

Daniel Ebbert
daniel.ebbert@adelaide.edu.au

Negin Mirriahi
negin.mirriahi@adelaide.edu.au

¹ Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan, Italy

² Centre for Change and Complexity in Learning, College of Education, Behavioural and Social Sciences, Adelaide University, Adelaide, SA 5000, Australia

³ School of Computer Science and Information Technology, College of Engineering and Information Technology, Adelaide University, Adelaide, SA 5000, Australia