Original Paper

Feasibility and Acceptability of AI-Generated Formative Assessments for Patient Immunomonitoring Training: Mixed Methods Implementation Study

Carlos Kiyan Tsunami[1], MD; Ermias Diro[2], MD, PhD; Odin Goovaerts[1], PhD; Aquiles Rodrigo Henríquez-Trujillo[1], MD; Karen Ferreira-Meyers[3,4], MA, MA, MIDT, LLM, PhD; Bart K M Jacobs[1], PhD; Ziyanda Mwanda[5], MDS, PGDHP; Jyotsna Rimal[6], BDS, MDS, PhD; Jamine Pozu-Franco[7], MEd; Héctor Javier Montalvo Navarrete[8], MD; Deogratias Katsuva Sibongwere[1], MD; Jean Michel Kolie[9], MSc; Gradi Luakanda-Ndelemo[10], MSc; Tchenagnon Claude Semevo[11], MSc; Sotheara Heng[12], MSc; Diljith Kannan[13], BSc; Harish Hn[14], BCA; Luis Fucay Guin[7], BA; Maria Zolfo[1], MD, PhD

[1]Institute of Tropical Medicine, Antwerp, Belgium
[2]University of Washington, Seattle, WA, United States
[3]University of Eswatini, Kwaluseni, Eswatini
[4]University of the Free State, Bloemfontein, South Africa
[5]School of Public Health, University of the Western Cape, Cape Town, South Africa
[6]University of Newcastle, Central Coast Campus, New South Wales, Australia
[7]Universidad Peruana Cayetano Heredia, Lima, Peru
[8]Pontificia Universidad Católica del Ecuador, Quito, Ecuador
[9]Centre National de Formation et de Recherche en Santé Rurale, Maferinyah, Guinea
[10]Institut National de Recherche Biomédicale, Kinshasa, Democratic Republic of the Congo
[11]Centre de Recherche en Reproduction Humaine et en Démographie, Cotonou, Benin
[12]National Institute of Public Health, Phnom Penh, Cambodia
[13]Indian Institute of Management Bangalore, Bangalore, India
[14]Institute of Public Health, Bangalore, India

Corresponding Author:

Carlos Kiyan Tsunami, MD
Institute of Tropical Medicine
Nationalestraat 155
Antwerp 2000
Belgium
Phone: +32 3 247 66 66
Email: ckiyan@itg.be

Abstract

Background: Traditional formative assessment development in specialized medical education is time-intensive and often poorly aligned with learning outcomes, particularly in highly technical domains such as patient immunomonitoring. Evidence on the feasibility and quality of AI-generated assessments in such specialized biomedical training remains limited.

Objective: This study aimed to evaluate the feasibility, implementation efficiency, expert review outcomes, and learner perceptions of a Claude 3.5 Sonnet–based approach for generating formative self-assessment questions from video-recorded lectures, contextualizing efficiency against lecturers' retrospective estimates of conventional manual question-writing time.

Methods: Using backward design and A-SMART (action-oriented, specific, measurable, achievable, relevant, and time-based) criteria, we processed 63 segments from 8 video-recorded lectures on cytokine profiling, flow cytometry, and immune-cell phenotyping. Claude 3.5 Sonnet (Anthropic), guided by specialized educational prompts, generated assessment questions from transcribed lecture content. Each question was independently reviewed by subject-matter experts for accuracy, relevance, cognitive-level appropriateness, and alignment with learning outcomes, and was then accepted without revision, revised with minor edits, or deleted. We tracked processing and review times and surveyed 12 course participants (70.6% response rate) on their perceptions.

Results: The large language model generated 439 assessment questions (7.0 per segment); 342 (77.9%) were accepted without edits, 37 (8.4%) after minor revision, and 60 (13.7%) were deleted, for an overall acceptance rate of 86.3% (n=379; 95% CI 82.8%-89.2%). Processing averaged 37 seconds per segment vs an estimated 4956 seconds for conventional manual question development, based on lecturers' retrospective estimates, corresponding to 99.3% time savings for generation alone and 84.4% when expert validation was included. Most questions were medium (n=251, 57.2%) or hard (n=165, 37.6%) difficulty, with medium-difficulty items showing the highest acceptance rate (228/251, 90.8%). Students rated helpfulness in understanding content and knowledge-gap identification (4.1/5 each); 83.3% (10/12) found the questions helpful for understanding content, and 75.0% (9/12) reported they helped identify knowledge gaps.

Conclusions: AI-assisted assessment generation from video lectures was feasible, was associated with substantial implementation-efficiency gains relative to lecturers' retrospective estimates, achieved high question retention following expert review, and was positively perceived by participating learners in a specialized medical education context. Future studies should determine whether this approach also improves learning outcomes, knowledge retention, or summative performance. This workflow offers a potentially transferable implementation model for formative assessment design, with relevance to health professional education across resource settings, pending evaluation in other disciplines and institutions.

JMIR Form Res 2026;10:e86309; doi: 10.2196/86309

Keywords: large language models; artificial intelligence; medical education; formative assessment; multimedia learning; patient immunomonitoring; educational technology; backward design; cognitive load theory; A-SMART learning outcomes

Introduction

Background

AI is reshaping education, introducing novel opportunities for automating instructional design and formative assessment [1-5]. In medical education, large language models (LLMs) such as ChatGPT offer potential for automated question generation, clinical-scenario creation, and personalized feedback to support self-directed learning [6,7].

Formative assessment, defined as assessment for learning rather than of learning, plays a critical role in promoting self-reflection, metacognition, and active learning [8-11]. Yet, in highly specialized domains such as patient immunomonitoring (PIM), developing valid and pedagogically aligned formative assessments is resource-intensive [12,13] and requires deliberate alignment with the intended learning outcomes.

Recent evidence confirms the growing feasibility of AI-based multiple-choice question (MCQ) generation in health professions education. A systematic review by Artsi et al [12] demonstrated that LLMs can produce competent questions suitable for medical examinations, with the process being fast and efficient, although accuracy and quality remain subject to expert oversight. More recently, Riehm et al [14] conducted a systematic review and network meta-analysis comparing AI-generated and human-authored MCQs across health professions education and found that ChatGPT-4–generated questions were of similar quality to human-authored items for relevance, clarity, and distractor quality, although the certainty of evidence was very low. Llama 2 performed worse than human authors on clarity and distractor quality, while exploratory comparisons found that ChatGPT-3.5 performed worse than Llama 2 and ChatGPT-4 on those outcomes. Most of this evidence concerns ChatGPT models; fewer published evaluations have examined other LLMs, such as Claude, for assessment generation in medical education. Despite this growing evidence base, few studies have examined AI-based assessment generation within highly specialized postgraduate biomedical training programs, and we are unaware of published studies that have integrated structured educational design frameworks such as backward design and A-SMART criteria within such a workflow.

Educational design frameworks, such as backward design [15] and constructive alignment [16], emphasize alignment between outcomes, learning activities, and assessment strategies. Cognitive load and multimedia learning theories provide complementary evidence that managing information load and segmenting complex material improves knowledge retention and comprehension [17-20].

To address these pedagogical and practical challenges, this study explores an AI-assisted approach for generating formative self-assessment questions directly from video-recorded lectures within an advanced PIM training program. The framework integrates A-SMART learning outcomes (action-oriented, specific, measurable, achievable, relevant, and time-based) with cognitive load and multimedia learning principles to maintain theoretical rigor while enhancing scalability and efficiency in assessment design [17,18,21].

This study examines the feasibility and acceptability of AI-generated formative assessments in terms of implementation efficiency and expert acceptance against prompt-defined quality standards, and offers a theory-driven implementation framework for formative assessment design in health professions education.

Theoretical Framework and AI Integration

Cognitive Load Theory and Multimedia Learning

The implementation of this study was guided by cognitive load theory [18,22] and by evidence that segmenting information in a way that reduces unnecessary cognitive effort improves learning [17,19,20]. Introduced by Sweller [22], this theory emphasizes that working memory has limited

capacity, and excessive cognitive load, especially from poorly structured material, can impede learning. Within the AI-based assessment generation framework, this theory informed both the segmentation of video-recorded lecture content into manageable units and the structuring of AI-generated assessment questions. The framework was intended to align each question set with a discrete segment of instructional content, consistent with cognitive load and multimedia learning principles.

The study also draws from multimedia learning principles, which recommend the integration of verbal and visual information and the segmentation of instructional content to improve understanding [17,23]. In this context, segmentation was applied to divide lectures into short, coherent segments (751-1250 words or approximately 6.5-8.5 min), so that each AI-generated question corresponded to a distinct learning element, thus reinforcing conceptual clarity and student engagement.

A-SMART Integration in Assessment Design

The study further applied the A-SMART framework as introduced by Kiyan Tsunami et al [21]. This framework operationalizes backward design principles [15] by aligning learning outcomes, instructional activities, and assessments within a coherent design cycle. Through A-SMART integration, each question generated by the LLM was mapped to a measurable and actionable learning objective, ensuring pedagogical integrity and alignment between instructional goals and formative testing.

The A-SMART framework was also instrumental in defining the prompt engineering parameters used for AI-assisted question generation. Prompts were constructed to specify difficulty level (easy, medium, or hard), content scope, and feedback format, building on established prompt engineering techniques [24] and reinforcing measurable outcomes while supporting student self-regulation. While the A-SMART framework supports the writing of clear and measurable learning outcomes, it does not, by itself, ensure assessment-item quality or define cognitive complexity criteria. The prompt requested easy/medium/hard difficulty labels but did not explicitly apply Bloom's taxonomy or provide criteria for classifying items by cognitive level. Consequently, difficulty classification was left to the model's interpretation of the transcript content.

Objectives

This study aimed to evaluate the feasibility and acceptability of a Claude 3.5 Sonnet–based framework for generating formative self-assessment questions from video-recorded PIM lectures. Specifically, it sought to (1) contextualize the time efficiency of AI-assisted assessment generation against lecturers' retrospective estimates of conventional manual question-writing time, (2) assess the quality of generated questions against prompt-defined standards through expert review, and (3) analyze student feedback regarding usability and perceived educational value. The overarching goal was to develop a theoretically grounded implementation framework and provide a foundation for future research on the transferability and educational effectiveness of AI-generated assessments in specialized medical education contexts.

Methods

Study Design

This was a single-course, mixed methods feasibility evaluation of an AI-assisted formative assessment workflow. Operational efficiency was contextualized against lecturers' retrospective, self-reported estimates of the time required for conventional manual item development; no concurrent human-authored comparator was evaluated. The evaluation also assessed question quality against prompt-defined standards and surveyed students' perceptions following implementation. All AI-generated questions were used exclusively for anonymous formative self-assessment purposes; no questions were used for grading, certification, summative evaluation, or any other high-stakes assessment purposes.

Data Sources and Learning Materials

Eight video-recorded lectures delivered during the 2024 (first) edition of the patient immunomonitoring 1 (PIM1) short course at the Institute of Tropical Medicine (ITM, Antwerp, Belgium) were used as data sources. PIM1 is the online preparatory phase of a 2-part international short course series: participants complete PIM1 remotely before attending the hands-on PIM2 component in Antwerp. The course is designed for clinical immunologists, particularly those involved in infectious diseases research in both high- and low-resource settings, and equips participants to design, implement, and interpret immunomonitoring studies to advance clinical and epidemiological research agendas. PIM1 was newly developed by the course coordinator, who holds a PhD in biomedical sciences with over 15 years of postdoctoral research experience. Instructional content was delivered by the course coordinator and a team of invited international lecturers, all at postdoctoral level or above, each with an established publication record in their respective content areas.

Each lecture covered advanced immunological topics, including cytokine profiling, flow cytometry, and immune-cell phenotyping. Lectures were recorded using either ScreenFlow (Telestream, macOS version) or Zoom (Zoom Video Communications), with the majority recorded using ScreenFlow. All recordings were edited using ScreenFlow prior to upload, including trimming and segment preparation. The edited recordings were then uploaded to a private YouTube course channel to facilitate automated transcription and accessibility. Lecture audio was transcribed using YouTube's automated captioning. Transcripts were used as generated, without manual correction, to simulate a realistic rapid-deployment scenario and assess the feasibility of the workflow under real-world conditions of imperfect automated transcription.

To optimize alignment with cognitive load theory and enable AI-based question generation, each of the 8 lectures was segmented into multiple shorter units representing coherent thematic sections. Segment boundaries were defined by topic shifts and natural pauses in instruction, resulting in segments of approximately 751 to 1250 words (about 6.5-8.5 min of speech). Each segment was subsequently paired with its corresponding learning outcomes and processed through the LLM for assessment generation.

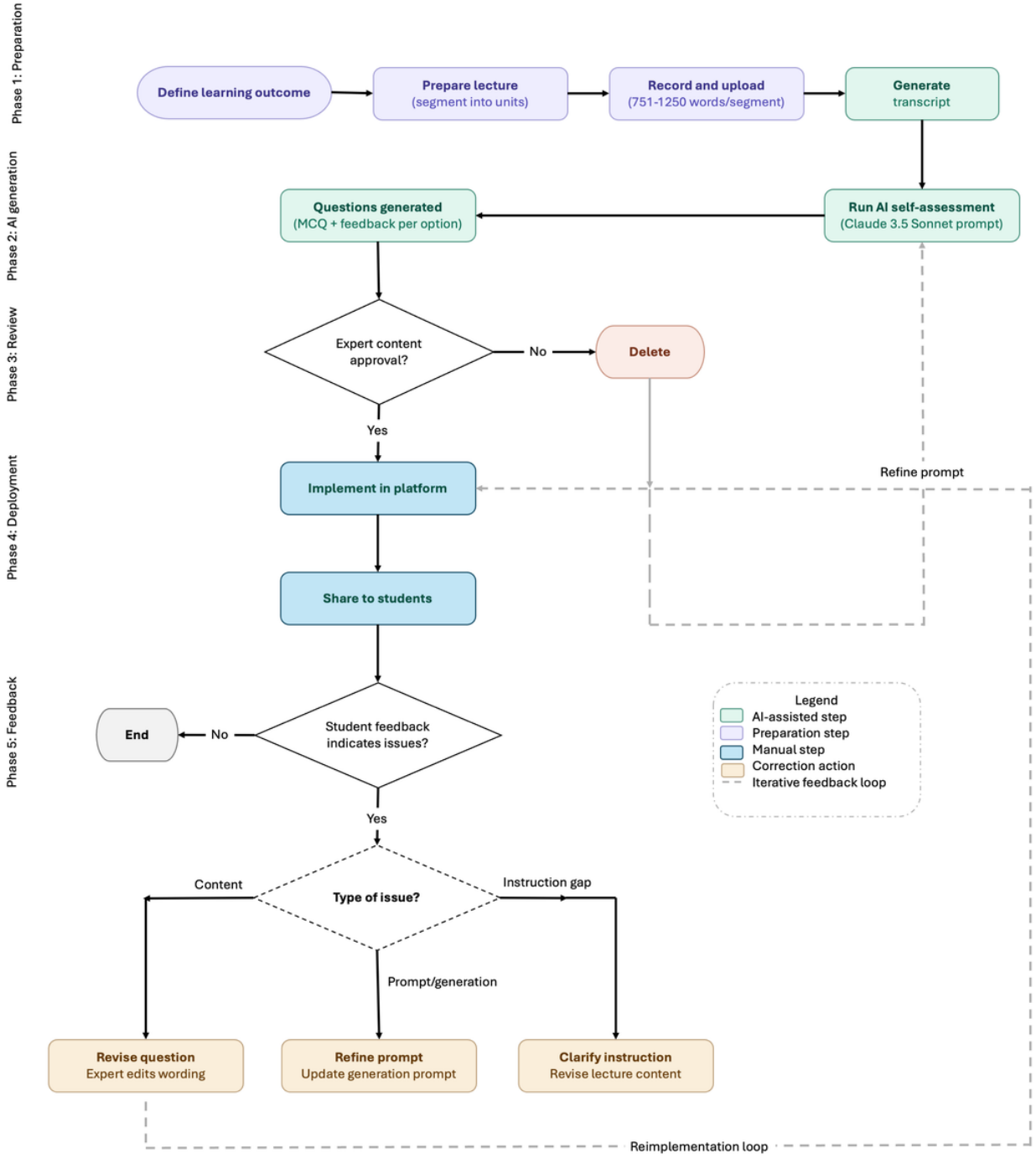
Framework Implementation

Backward Design and Lecture Segmentation

The implementation followed 3 core elements, drawing on established educational design principles. This study used a backward design approach with preplanned lecture segmentation. Before recording, subject-matter experts (SMEs) were instructed to divide their lectures into thematic segments of approximately 751 to 1250 words each to optimize cognitive load and create manageable AI-processing units. In practice, segment lengths varied across lectures, as not all instructors fully adhered to this guideline, resulting in a total of 63 segments of varying lengths. Each segment was intended to address discrete, measurable learning objectives aligned with the course design. The relationship between segment length and AI-processing efficiency is reported in the Results section, and because most segment-length categories contained few segments, this analysis is presented descriptively rather than as evidence that any specific range is optimal.

After recording, the prompt framework was applied to each transcript segment in 2 sequential steps. In the first step (part 1 of the prompt), Claude 3.5 Sonnet regenerated learning outcomes directly from the transcript content following A-SMART criteria, assigning unique learning outcome (LO) identifiers (eg, LO.01 and LO.02). The AI-regenerated outcomes were available for comparison with the SME-defined outcomes as a workflow-level alignment check between instructional intention and delivered content; concordance between the 2 sets was not formally analyzed as a study outcome. In the second step (part 2 of the prompt), Claude 3.5 Sonnet generated 2 assessment questions per regenerated LO, ensuring that all questions were grounded in what was actually taught. The course content covered key immunological topics, including cytokine profiling, flow cytometry, and immune-cell phenotyping. The framework consisted of 3 steps (Figure 1).

Figure 1. Workflow used for AI-assisted assessment generation. A schematic overview of the workflow used for AI-assisted assessment generation shows the integration of lecture segmentation, A-SMART–aligned prompts, and expert review for quality assurance. A-SMART: action-oriented, specific, measurable, achievable, relevant, and time-based; MCQ: multiple-choice question.



Content Preparation

Lecture content was segmented into shorter units, each mapped to specific learning outcomes. Segment lengths varied across lectures, and the relationship between segment length and processing efficiency is reported in the Results section. Transcripts were used directly from YouTube's automated captioning without preprocessing, reflecting a deliberate choice to evaluate the workflow under realistic rapid-deployment conditions.

AI Processing

This step involves automated question generation using tailored prompts aligned with A-SMART criteria. The complete prompt structure, including formatting requirements, difficulty-distribution guidelines, and feedback instructions, is available in Multimedia Appendix 1.

Human Review

SMEs validated and refined the questions for accuracy, alignment, and adherence to prompt-defined quality standards. This step maintained educational integrity while leveraging AI efficiency.

AI-Assisted Question-Generation Process

Following the A-SMART framework [21] and principles of backward design [15], each transcript segment was used as input for the Claude 3.5 Sonnet–based question-generation process (Anthropic, October 2024 release; accessed via the claude.ai web interface using default model settings). Prior to full deployment, 2 lecture samples were independently processed by both Claude 3.5 Sonnet and GPT-4o (OpenAI, November 2024; accessed via the ChatGPT web interface using default model settings) to inform model selection. Two SMEs reviewed the outputs and reached consensus that Claude 3.5 Sonnet produced superior question formulation and pedagogical clarity. Accordingly, Claude 3.5 Sonnet was selected as the sole model for all subsequent question generation, and all 439 questions in the final dataset were generated exclusively using this model. A custom educational-design prompt library was developed to standardize the structure and output of the model across all segments.

Claude 3.5 Sonnet was prompted to generate MCQs aligned with each content segment. Each question offered 4 or 5 answer options (A-E, with a fifth option used where appropriate), with a single correct option for MCQs and multiple correct options for multiple-response (MR) items. The prompt targeted 90% single-answer MCQs and 10% MR items (Multimedia Appendix 1). Item format was not recorded separately in the analytic dataset; therefore, all review outcomes are reported at the item level, and the retrospective manual-time benchmark, elicited for MCQs, should be interpreted as an approximate benchmark for the overall workflow. Difficulty levels (easy, medium, or hard) were initially assigned by Claude 3.5 Sonnet in response to difficulty labels requested in the prompt framework. Because the prompt did not provide an explicit taxonomy or predefined criteria for difficulty classification, these assignments relied on the model's interpretation of the transcript content. During expert review, questions were assessed for appropriateness of difficulty relative to the intended LOs; questions deemed too simple or insufficiently challenging for the target audience were deleted. The prompt's guideline favoring medium and hard items (Multimedia Appendix 1) likely explains the low proportion of easy questions (n=23, 5.2%) among the 439 generated questions; questions deemed too simple were additionally deleted during expert review. No formal psychometric analysis was conducted to independently validate the AI-assigned difficulty classifications. The position of the correct answer was systematically varied across questions to minimize response bias, for example, ensuring that the correct answer did not consistently appear in the same position (eg, always option A or always option C) across consecutive questions. This variation was embedded directly in the prompt framework, which instructed Claude 3.5 Sonnet to distribute correct answer positions across all available options throughout the question set. The model also produced individualized formative feedback for each answer choice, explaining the rationale for both correct and incorrect responses to support self-regulated learning.

All AI-generated assessment questions were reviewed for adherence to the predefined item-writing standards specified in the prompt framework, including content accuracy, distractor plausibility, cognitive-level appropriateness, feedback quality, and technical clarity. Minor grammatical and contextual edits were performed when necessary, but the semantic content and feedback wording were retained to preserve the model's original pedagogical framing.

Expert Validation and Quality Review

The review process was not blinded; all reviewers were aware that the questions had been generated by Claude 3.5 Sonnet prior to conducting their evaluations. The expert review process involved 2 distinct roles. The first author oversaw prompt engineering and ensured that all generated questions adhered to established single-answer and MR item-writing standards by specifically verifying the absence of guessing cues, implausible distractors, and structural flaws that could compromise item validity.

Content validation was conducted by the course coordinator and the SME responsible for each respective lecture. Each SME reviewed the questions generated from their respective lecture segments for factual accuracy, conceptual relevance, clarity, and cognitive-level alignment with the intended LOs. The course coordinator reviewed all questions across all lectures and made the final decision regarding acceptance, minor revision, or deletion of each item.

The SME and the course coordinator reviewed each generated question independently, recording their own accept/edit/delete decision without access to the other reviewer's judgment, before these independent decisions were compared. When differences in judgment arose between the SME and the course coordinator, the course coordinator's decision was applied as the final adjudication rule.

Question quality was assessed against the predefined item-writing standards embedded in the prompt framework (Multimedia Appendix 1), which specified criteria across 5 dimensions: (1) content alignment: direct relevance to lecture content and explicit mapping to specific learning outcomes; (2) distractor quality: plausible distractors of similar length, homogeneous categories, absence of obvious guessing cues, and no use of "all of the above" options; (3) cognitive-level appropriateness: the appropriateness of the AI-assigned difficulty level (easy, medium, or hard) in relation to the question complexity and targeted learning outcome (operationalized for experts using the same prompt-embedded criteria described above, namely, whether the cognitive demand implied by the question matched the depth of understanding specified in the corresponding learning outcome, rather than an independently validated psychometric taxonomy such as Bloom's taxonomy); (4) feedback quality: individualized explanatory feedback for every answer option,

explicitly labeled as correct or incorrect with a complete explanatory sentence; and (5) technical quality: this included clear and unambiguous language, complete question stems, and proper grammar. While these criteria were systematically embedded in the generation prompt and applied consistently during review, no separately validated psychometric rubric was used. Expert review therefore reflected structured professional judgment against predefined item-writing standards embedded in the prompt framework, covering content alignment, distractor quality, appropriateness of the AI-assigned difficulty level, feedback quality, and technical clarity.

For the purposes of this study, the following operational definitions were applied consistently throughout the review process:

- Accepted (no edits): questions that passed expert review exactly as generated, requiring no modifications
- Edited (minor edits): questions that required minor corrections, such as grammatical adjustments, rewording for clarity, or contextual refinements, before being approved for use
- Deleted: questions that were completely removed due to factual inaccuracies, conceptual ambiguities, or misalignment with the intended learning objectives, and were deemed unsalvageable through minor editing.
- Accepted-without-edits rate: the proportion of AI-generated questions accepted without edits, representing questions considered immediately usable by experts without modification.
- Acceptance rate: the proportion of questions retained for use, including both those accepted without edits and those accepted after minor editing, relative to the total number of questions generated (accepted without edits + edited) divided by total questions generated.

Expert review time was recorded at the lecture level for 7 lecture/lecturer blocks and was used to derive segment-level review-time estimates for the 47 segments with available timing data, which were then used to calculate relative efficiency compared with manual development.

Student Survey and Feedback Collection

Following completion of the AI-supported 2024 PIM1 course, all 17 enrolled participants were invited to complete an anonymous postcourse survey assessing usability, perceived learning support, and satisfaction with the AI-assisted assessments. The survey instrument, consistent with established principles of questionnaire development (see Koo and Yang [25] for a recent summary), comprised 2 parts: (1) 12 general items covering overall helpfulness, knowledge-gap identification, content match, question quantity and difficulty, clarity, feedback usefulness, confidence, and preparation for formal evaluations, including open-ended items analyzed qualitatively (see Data Analysis section); and (2) module-specific ratings across 16 course chapters. General survey items used 5-point Likert-scale or categorical response formats. Module-specific ratings included 5-point helpfulness ratings, 3-category difficulty ratings, and binary yes/no items. Twelve students responded, yielding a response rate of 70.6% (12/17), calculated and reported in accordance with established response rate reporting guidelines [26]. The complete survey instrument is provided in Multimedia Appendix 2.

Data Analysis

Quantitative data were analyzed descriptively using counts, percentages, means, and SDs. Comparative efficiency between AI-assisted assessment generation and conventional manual question development was calculated as the percentage reduction in total time required. Expert evaluation outcomes were summarized as frequency distributions across 3 mutually exclusive categories (accepted without edits, accepted after minor edits, and deleted). Survey responses were summarized using descriptive statistics for Likert-scale items and frequency distributions for categorical and binary items. Open-ended responses to items 9 and 10, asking, respectively, about aspects of the system that most helped learning and suggested improvements, were examined using an inductive descriptive approach: the first author reviewed all responses and grouped similar comments into recurring descriptive categories, without a predefined coding framework or codebook. Given the small number of open-ended responses (9/12 and 10/12 respondents, respectively), this approach was considered more appropriate than formal thematic analysis. This coding was performed by a single reviewer (the first author) without a second independent coder or a formal intercoder reliability check; this is acknowledged as a limitation below.

The processing efficiency index (PEI) was calculated for each transcript-length category using the following formula:

PEI = average processing speed (words processed per minute) ÷ revision-or-deletion rate (items requiring revision or deletion per segment). Processing speed was calculated as each segment's transcript (input) word count divided by Claude 3.5 Sonnet's recorded response time for that segment, then averaged within each word-count category. A "revision or deletion" was defined as any generated question that required expert editing or deletion during review (ie, any outcome other than acceptance without edits); edits and deletions were weighted equally and summed to give a per-segment count, which was then averaged across all segments within each word-count category to derive the category-level rate used in the PEI denominator.

Higher PEI values indicate greater processing efficiency by combining faster processing with fewer items requiring revision or deletion. PEI was developed specifically for this implementation study as a descriptive metric for comparing transcript segment lengths and is not intended as a standardized or psychometrically validated performance measure.

To establish a baseline for time-efficiency comparison, a brief survey was administered to 6 course lecturers (A, B, C, D, F, and H) involved in the PIM program. Each lecturer was asked to estimate the minimum and maximum time required to create and review one MCQ of comparable complexity to those generated by the AI framework, from scratch and based on their area of expertise. The survey provided a reference example showing the full question, answer options, correct

answer identification, individualized explanatory feedback for every answer option, learning outcome assignment, and difficulty level classification. The estimate, therefore, covered the complete question development and self-review process. LMS formatting and upload time were not included in the estimate. The resulting estimates were used as the baseline for manual question development time in all subsequent efficiency calculations and are reported in the Results section.

Ethical Considerations

This study received ethical approval from the institutional review board of the Institute of Tropical Medicine, Antwerp (protocol ref: 1842/25; approval date: February 20, 2025). All participants provided written informed consent before completing the feedback survey. Participants did not receive any financial or other compensation for completing the survey. Data were anonymized and managed in accordance with institutional privacy and data protection policies.

This study did not involve the prospective assignment of human participants to a health-related intervention and, therefore, does not meet ICMJE criteria for a registerable clinical trial.

Results

AI-Generated Question Characteristics

Claude 3.5 Sonnet generated a total of 439 assessment questions from 63 segmented lecture units across 8 video-recorded lectures (Table 1). Each segment produced an average of 7.0 questions, aligned with its corresponding learning outcomes. Following expert review, 86.3% (n=379; 95% CI 82.8%-89.2%) of the AI-generated assessment questions were accepted for formative use (acceptance rate). Of all generated questions, 77.9% (n=342; 95% CI 73.8%-81.5%) were accepted without edits (accepted-without-edits rate), while 8.4% (n=37) were accepted after minor revisions. The remaining 13.7% (n=60) were deleted due to conceptual or structural issues.

Table 1. Expert review outcomes of AI-generated assessment questions across patient immunomonitoring lectures.

Lecture	Lecturer	Review time	Segments	Minor edits[a]	Deleted[b]	Total questions	Accepted without edits, n (%)[c]	Acceptance rate, n (%)[d]
1	A	01:20:00	10	13	16	75	46 (61.3)	59 (78.7)
1	H	00:40:00	2	4	2	20	14 (70.0)	18 (90.0)
2	B	00:59:00	4	2	9	30	19 (63.3)	21 (70.0)
3	C	01:32:00	8	0	7	46	39 (84.8)	39 (84.8)
4	A	01:15:00	11	9	5	72	58 (80.6)	67 (93.1)
5[e]	D	00:26:00	3	8	0	12	4 (33.3)	12 (100)
6	E	—[f]	7	1	4	44	39 (88.6)	40 (90.9)
7	F	03:24:00	9	0	17	100	83 (83.0)	83 (83.0)
8	G	—	9	0	0	40	40 (100)	40 (100)
Total[g]	—	09:36:00	63	37	60	439	342 (77.9)	379 (86.3)

[a]Minor edits=questions requiring minor corrections before approval.
[b]Deleted=questions removed due to inaccuracy, ambiguity, or learning-outcome misalignment.
[c]Accepted-without-edits rate=questions accepted without edits/total questions generated.
[d]Acceptance rate = (questions accepted without edits + questions accepted after minor edits) / total questions generated.
[e]Lecture 5 shows an accepted-without-edits rate of 33.3% with an acceptance rate of 100% because all 12 generated questions were retained for use (0 deleted); however, 8 of 12 questions required minor editing before approval, resulting in only 4 questions (33.3%) accepted without edits.
[f]Not available.
[g]Total recorded review time was 34,560 seconds (09:36:00) across 47 segments from 7 lecture/lecturer blocks. The mean review time was 735.3 (SD 361.5) seconds (12.3 min) per timed segment; review-time data were unavailable for lectures 6 and 8. Segment-level calculations are reported in Multimedia Appendix 3.

To illustrate the range of expert review decisions, 3 representative examples are summarized here; full annotated examples are provided in Multimedia Appendix 4. First, an accepted question (no edits required) asked students to identify the ideal droplet composition for generating a single-cell library in a 10× Genomics workflow (medium difficulty; LO.02). This question was accepted without modification owing to its unambiguous stem, 4 plausible and similarly structured distractors free of guessing cues, and individualized explanatory feedback for every answer option. Second, a question requiring minor edits asked students to identify the most appropriate analytical approach for working with whole-blood bulk RNA sequencing data (hard difficulty; LO.02). The original AI-generated stem was ambiguous in its intent; the expert added the phrase “to study T cell subset contributions” to clarify the analytical context, while leaving the answer options, correct answer, and feedback unchanged. Third, a deleted question asked students to identify the most appropriate experimental setup for a carboxyfluorescein succinimidyl ester proliferation assay to evaluate an immunosuppressive drug (hard difficulty; LO.01). Despite being technically accurate, this question was deleted because it tested recall

of specific procedural steps rather than comprehension of the underlying immunological principles targeted by the learning outcome and because the correct answer was highly context-dependent in ways not fully captured by the available options. These examples illustrate that expert review decisions were grounded in pedagogical appropriateness and learning-outcome alignment, not solely in factual accuracy.

Efficiency and Expert Validation Outcomes

Question Difficulty Analysis

Analysis of question difficulty revealed that the majority of AI-generated questions were classified as medium difficulty (251/439, 57.2%), followed by hard (165/439, 37.6%) and easy (23/439, 5.2%) categories (Table 2). Medium difficulty questions achieved the highest acceptance rate (90.8%) and accepted-without-edits rate (82.5%), whereas easy questions showed lower acceptance rates, although this finding should be interpreted cautiously given the small number of easy items (n=23).

Table 2. Distribution of questions by difficulty level and editing requirements.

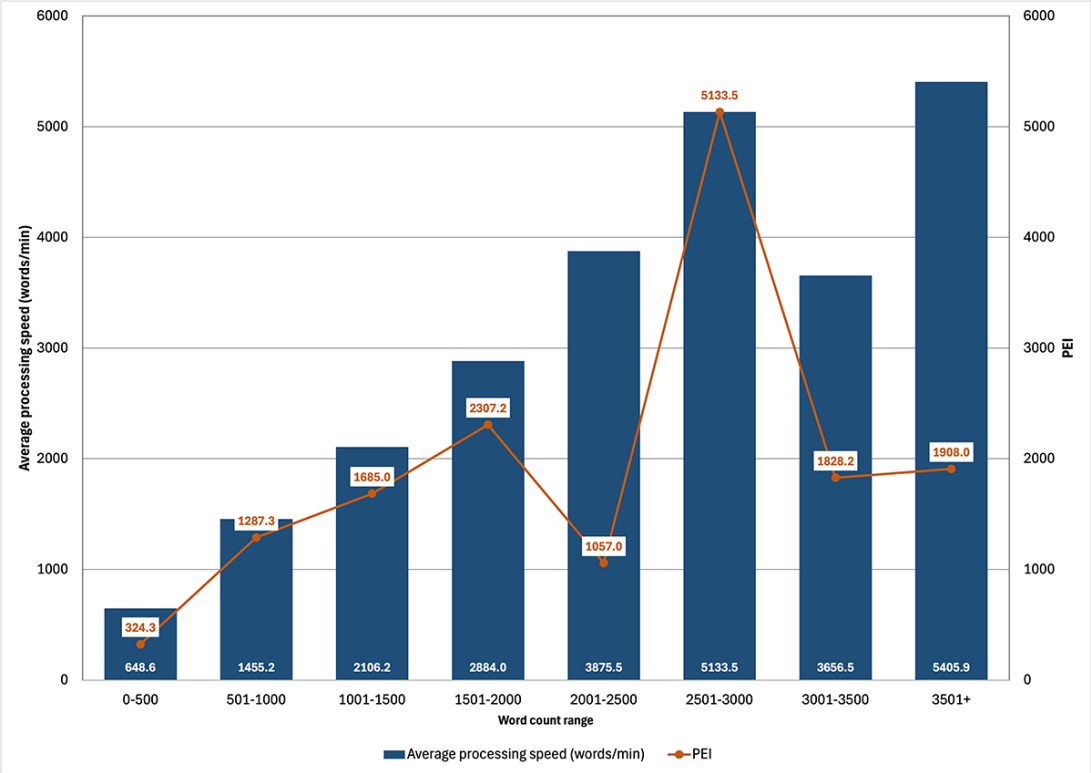
Difficulty level	Total questions	No edits[a]	Minor edits[b]	Deleted[c]	Accepted without edits, n (%)[d]	Acceptance rate[e], n (%)	Distribution, n/N (%)
Easy	23	14	2	7	14 (60.9)	16 (69.6)	23/439 (5.2)
Medium	251	207	21	23	207 (82.5)	228 (90.8)	251/439 (57.2)
Hard	165	121	14	30	121 (73.3)	135 (81.8)	165/439 (37.6)
Total	439	342	37	60	342 (77.9)	379 (86.3)	439/439 (100)

[a]No edits indicate questions accepted exactly as generated.
[b]Minor edits indicate questions requiring minor corrections before approval.
[c]Deleted means questions removed due to inaccuracy, ambiguity, or misalignment.
[d]Accepted-without-edits rate = no edits / total questions.
[e]Acceptance rate = (no edits + minor edits) / total questions.

Segment Efficiency by Word Count

Analysis of processing efficiency by segment length is presented descriptively (Multimedia Appendix 3, Table S1, Figure 2), using 500-word categories; a finer 250-word categorization was examined initially but was not used for reporting, as several of its narrower categories contained only one or two segments, producing PEI values driven by individual-segment variability rather than a stable estimate. Among the 3 categories with the largest samples, 501 to 1000 words (23 segments), 1001 to 1500 words (16 segments), and 1501 to 2000 words (8 segments), average processing speed and PEI increased consistently with segment length: 1455.2 words/minute (PEI 1287.3), 2106.2 words/minute (PEI 1685.0), and 2884.0 words/minute (PEI 2307.2), respectively. This pattern, in which processing speed increases with segment length, may reflect a fixed per-call processing overhead that is amortized over more words as segment length increases and should not be interpreted as evidence that shorter segments are processed more efficiently. The remaining categories contained few segments (as few as one) and produced more variable PEI values that likely reflect small-sample variability rather than genuine differences in efficiency; segment length was also confounded with lecture topic, lecturer, and transcript quality. The segmentation range of 751 to 1250 words used in this study was chosen for pedagogical reasons consistent with cognitive load theory and prior evidence that students often watch less than half of instructional videos longer than 9 minutes [27,28], rather than because it was empirically optimal for AI processing efficiency.

Figure 2. Processing efficiency by transcript word count range. Bars represent average processing speed (transcript word count divided by Claude 3.5 Sonnet's response time for that segment, in words per minute), while the line represents the processing efficiency index (PEI), calculated as the average processing speed divided by the revision-or-deletion rate (items requiring revision or deletion per segment). Higher PEI values represent greater processing efficiency through the combination of faster processing and fewer items requiring revision or deletion. Word count categories are defined in 500-word increments. The 3 largest categories, 501 to 1000 words (23 segments), 1001 to 1500 words (16 segments), and 1501 to 2000 words (8 segments), provide the most reliable basis for comparison and show PEI increasing with segment length (1287.3, 1685.0, 2307.2, respectively). The remaining categories contain few segments, so their PEI values should be interpreted with caution.



Processing Time Metrics

The LLM demonstrated consistent and rapid processing performance across all lecture segments. The mean response time per segment was 37 seconds, with a range from 24 to 70 seconds (Multimedia Appendix 3, Table S2). The 90th percentile response time was 51 seconds, and the SD was 10 seconds, showing relatively limited variability across inputs.

By contrast, SME review times were longer and more variable (Multimedia Appendix 3, Table S3). The average review time per segment, derived from total lecture-level review times, was 735.3 (SD 361.5) seconds (12.3 min), with a median of 520 seconds (8.7 min). The shortest review took 409.1 seconds (6.8 min), while the longest required 1360 seconds (22.7 min). The IQR of 480 to 885 seconds indicates moderate variability, reflecting differences in segment complexity and the degree of manual refinement required to finalize each question set.

Time Efficiency Comparison

Comparison with conventional manual question development revealed substantial time savings across all efficiency scenarios. Based on a survey of 6 course lecturers, manual development of one MCQ of comparable complexity, including question stem, distractors, and explanatory feedback, required an average of 11.8 (SD 3.9; median 11.3, IQR 8.5-15.0) minutes per question; this estimate was based on the MCQ format specifically, and an equivalent manual-time benchmark for the smaller proportion of MR items was not separately collected. Each lecturer's own average ranged from 7.5 to 17.5 minutes, while individual per-question estimates across all 6 lecturers ranged more widely, from 5 to 20 minutes. Given that each segment produced an average of 7.0 questions, manual creation would take approximately 4956 seconds (82.6 min) per segment.

Using Claude 3.5 Sonnet, the initial automated question generation time was reduced by 99.3% (37 seconds vs 4956 seconds). However, as expert review is an essential component of the workflow, the more practically meaningful efficiency measure is the total processing time including human validation. When incorporating the mean expert review time of 735.3 seconds per segment, the total AI-assisted processing time was 772.3 seconds per segment, representing an overall time saving of 84.4% relative to lecturers' retrospective estimates of conventional manual question development. This figure, representing a reduction from 82.6 minutes to approximately 12.9 minutes per segment, is considered the primary efficiency estimate for this study, as it reflects the complete operational workflow, including quality assurance.

Using the highest lecturer-average estimate of 17.5 minutes per question, the estimated time saving including expert review was 89.5%. Using the lowest estimate of 7.5 minutes per question—the more conservative assumption

—the corresponding saving was 75.5%. These sensitivity analyses suggest that the estimated time saving remained substantial, above 75%, across the range of lecturer-average estimates.

Student Feedback and Perceptions

A total of 12 out of 17 course participants completed the post-course survey, yielding a 70.6% response rate. Overall, students expressed positive perceptions of the AI-generated formative assessments (Table 3). The highest ratings were observed for helpfulness in understanding content and knowledge-gap identification (both 4.1/5), followed by relevance to learning outcomes (3.9/5), feedback usefulness (3.7/5), and question clarity (3.8/5). All 12 respondents (100%) rated the difficulty level of the questions as appropriate ("just right"), with none rating them as too easy or too difficult.

Table 3. Student ratings and perceptions of assessment characteristics.

Assessment aspect	Average rating
Helpfulness in understanding content	4.1
Relevance to learning outcomes	3.9
Knowledge-gap identification	4.1
Question clarity	3.8
Appropriate difficulty level	100% "Just right"
Feedback usefulness	3.7

Students rated the self-assessment questions positively overall, with 83.3% (10/12) finding them helpful or extremely helpful in understanding lecture content. Most respondents (9/12, 75%) reported that the questions helped identify gaps in their understanding, and 75% (9/12) rated the questions as well-matched to the lecture content. All 12 respondents (100%) found the difficulty level appropriate, and 91.7% (11/12) considered the number of questions per segment appropriate.

Open-ended responses (9/12 respondents for item 9; 10/12 for item 10) were examined using an inductive descriptive approach and yielded 3 descriptive categories related to perceived benefits and 3 related to suggested improvements.

Regarding perceived benefits, 3 categories were identified: (1) active recall and content reinforcement, with students reporting that the rhythm of questions following each lecture segment encouraged active engagement and helped consolidate lecture content; (2) formative feedback mechanisms, with participants valuing the explanations provided for correct and incorrect answer options and noting that these supported self-directed learning and comprehension; and (3) content alignment, with students describing the questions as closely aligned with lecture content and helpful for identifying gaps in understanding.

Regarding suggested improvements, 3 categories emerged: (1) content accuracy concerns, with multiple respondents reporting instances in which correct answers appeared to be marked as incorrect; subsequent review determined that these discrepancies resulted from technical errors during the copy-and-paste transfer of questions to the self-assessment platform rather than from inaccuracies in the AI-generated content itself; (2) question bank diversity, with students suggesting the development of larger rotating question banks to reduce repetition across multiple attempts; and (3) alignment with formal assessment, with some respondents indicating that the self-assessment content did not always reflect the emphasis of the final examination and that highly specific questions concerning numerical values or details from individual papers had limited practical value.

These findings should be interpreted with caution; however, given the small absolute sample size (n=12), the single institutional setting, and the voluntary nature of survey participation, which may have introduced response bias, with students who found the AI-generated assessments more helpful potentially more likely to respond. The modest sample size also limits statistical power and generalizability to other learner populations or educational contexts.

Discussion

Principal Results

This study demonstrated the feasibility of integrating Claude 3.5 Sonnet into assessment design for the PIM course, achieving substantial efficiency gains, high item retention following expert review, and positive student perceptions. Acceptance reflects expert judgment against predefined standards and should not be interpreted as an objective psychometric measure of item quality. Open-text responses described perceived benefits related to active recall, formative feedback, and content alignment. The LLM produced 439 assessment questions from 63 lecture segments, reducing total assessment development time by 84.4% relative to lecturers' retrospective estimates of conventional manual question development. More than 86% of the generated questions were accepted by experts with minimal edits, and student feedback indicated positive perceptions of helpfulness, content alignment, and knowledge-gap identification. Together, these findings support the feasibility of an AI-assisted assessment generation framework in which AI enhances efficiency while structured expert review remains essential to ensure content accuracy, pedagogical appropriateness, and alignment with intended learning outcomes.

Comparison With Prior Work

The segmentation design drew on prior evidence that structured, segmented content enhances student attention and reduces cognitive overload [17,20,28]. The 6.5 to 8.5-minute segment duration used in this study was selected for pedagogical reasons, informed by evidence that engagement with instructional video declines as length increases [27]; the segments used here exceeded the less than 6-minute length recommended by Guo et al [27], and learner cognitive load was not directly measured.

In line with Wiggins and McTighe's Understanding by Design framework [15] and the A-SMART learning outcome model [21], the AI-assisted workflow was designed to maintain backward alignment between learning objectives and formative assessment. This reflects principles of constructive alignment and formative feedback described in the literature [11,16].

The student feedback findings are consistent with emerging evidence that AI-generated practice questions and explanations are viewed positively by learners and faculty as supplemental learning tools. In a faculty-evaluated study, Tong et al [29] found that ChatGPT shows promise for generating practice MCQ explanations for medical students; Elzayyat et al [30] found that medical students rated AI-generated anatomy MCQs as closely approaching expert-written items; and Zhang et al [31] demonstrated the feasibility and positive student reception of AI-generated clinical cases with MCQs in undergraduate medical education. Taken together, these studies suggest that AI-generated practice questions and explanations are generally perceived positively by learners and faculty across a range of medical education contexts, while underscoring that expert oversight remains an essential component of quality assurance. This principle is embedded in the AI-assisted assessment generation framework evaluated in the present study.

Several additional studies contextualize these findings. Arif et al [32] demonstrated the feasibility of generating MCQs directly from video lecture transcripts using LLMs, a workflow that closely parallels the approach described in the present study, and highlighted the importance of structured evaluation frameworks for automated question generation. Law et al [13] found similar item discrimination (although AI-generated items were easier) and substantial time savings for ChatGPT-4o–generated vs human-authored MCQs among candidates preparing for a high-stakes emergency medicine specialty examination, consistent with the efficiency gains observed in the present study. Most of these comparison studies evaluated ChatGPT models; in the present study, Claude 3.5 Sonnet was selected after a pilot comparison with GPT-4o (see Methods section), and whether these findings transfer to other LLMs remains to be tested. Camarata et al [33] evaluated AI-generated practice quizzes for preclinical medical students and identified item-writing flaws, including inappropriate distractors, pseudocase vignettes, and factual or conceptual errors, reinforcing the need for expert review before deployment. Collectively, these studies support the conclusion that AI-based MCQ generation is emerging as a feasible and efficient approach across medical education contexts, with expert oversight consistently identified as essential for quality assurance.

From an AI-education perspective, our results reinforce recent findings that generative models can produce usable question sets [32] and explanations [29] and substantially reduce instructor workload [13]. The observed efficiency gains (84.4% time savings, including expert validation) are consistent with the time savings reported by Law et al [13] for the overall creation of AI-generated vs human-generated MCQs (24.5 vs 96 person-hours); their magnitude may reflect automated segmentation and streamlined review integration.

Practical Implications

It is important to interpret the reported time savings in their appropriate context. While the 99.3% reduction in automated generation time is technically accurate, this figure excludes the essential expert review step and should not be considered the primary measure of efficiency. The more practically meaningful estimate, 84.4% time savings including expert validation, reflects the complete operational workflow and is therefore the figure most relevant to institutions considering adoption of this framework. Even under the most conservative assumptions based on the range of lecturer-average estimates, time savings including expert review remained above 75%.

By demonstrating substantial time savings and positive student reception, this approach illustrates one implementation pathway for integrating AI into health professional education, though its transferability to other courses, institutions, and resource settings remains to be evaluated. The method can support instructors in generating formative assessments aligned with predefined outcomes, while maintaining educational rigor through expert validation. In low-resource or time-constrained contexts, the model may lower the time required for assessment development relative to fully manual approaches, subject to local validation and workflow costs, though expert review remains an essential and nontrivial component of the workflow.

Limitations

Several limitations should be considered. In terms of scope and outcomes, the study was conducted in a single specialized course (PIM) with a modest, voluntary sample of 12 student respondents, which constrains statistical power and generalizability to other learner populations, course contexts, institutional settings, or subject areas. The voluntary nature of participation may have introduced response bias, with students who found the assessments more helpful potentially more likely to respond, so student-perception findings should be interpreted as preliminary and descriptive. No LO or psychometric-effectiveness comparison was conducted; the reported acceptance and learner-perception findings reflect feasibility and acceptability rather than evidence of improved learning.

In terms of measurement and comparison, the baseline for manual question-development time relied on 6 lecturers' retrospective self-report estimates rather than directly timed observations, and review-time data were incomplete for 2 of 8 lectures; future studies could refine this comparison using controlled, directly timed manual question-writing tasks and complete review-time capture. The 37-second AI processing time reflects only Claude 3.5 Sonnet's per-segment response time and does not include the upfront time invested in designing and refining the prompt framework, which took approximately a few hours and was a one-time cost amortized across all 63 segments; this was not formally tracked and is not reflected in the reported efficiency estimates. Expert review followed a structured, prompt-embedded rubric rather than a separately validated psychometric instrument; no formal interrater reliability statistic (eg, Cohen κ) was calculated, and the qualitative analysis of open-ended survey responses was conducted by a single reviewer without a second coder; future work could incorporate formal reliability statistics for both. Members of the implementation team, including the first author who oversaw prompt engineering, also participated in expert review and qualitative coding activities; while review decisions were recorded independently for each reviewer, this overlap between development and evaluation roles is common in implementation studies but could introduce confirmation bias.

In terms of technology and implementation, the AI model's output quality depends on prompt specificity and may vary across content domains or AI versions; because Claude 3.5 Sonnet represents a single snapshot of a rapidly evolving technology, the reported acceptance and accepted-without-edits rates should not be assumed to generalize to earlier or later model versions without independent re-evaluation. The study also relied on YouTube's automated transcription, which may have introduced transcription inaccuracies. AI-based assessment generation carries AI-specific risks, including hallucinated content, subtly flawed distractors, unintended answer cueing, misalignment with intended learning outcomes, and inconsistent difficulty calibration; these risks reinforce the essential role of structured expert review before deployment, and in this study all AI-generated questions were used exclusively for formative self-assessment, with none deployed for summative evaluation.

Conclusions

Future work should investigate how AI-generated assessments influence actual learning outcomes and performance metrics in health professional training. Expanding to diverse topics and participant populations could validate the model's generalizability. Question difficulty in this study was assigned by Claude 3.5 Sonnet and filtered through expert judgment rather than formal psychometric validation (eg, item difficulty or discrimination indices); future studies could incorporate such analyses and integrate adaptive algorithms for automatic difficulty calibration to improve personalization and formative feedback.

In conclusion, the AI-assisted workflow described here appears feasible, efficient, and acceptable to learners in this specific course context. These findings support the potential utility of this approach as a time-saving tool for formative assessment design in specialized medical education, provided that structured expert oversight is maintained as an essential component of the workflow. Whether this approach translates into measurable improvements in objective learning outcomes, knowledge retention, or summative performance remains an important question requiring further study. When combined with established educational design principles such as cognitive load optimization and A-SMART alignment, this workflow offers a theoretically grounded foundation for future research on AI-assisted assessment design in health professional education.

Acknowledgments

The authors wish to thank the ITM, Antwerp, Belgium, for providing institutional and technical support for this study and for granting access to course materials from the patient immunomonitoring program. The authors also express their gratitude to the course participants for their valuable feedback and engagement during the evaluation process.

The authors acknowledge the contribution of faculty reviewers and academic coordinators who participated in the expert validation process and ensured the educational rigor of the assessment items.

Generative AI Use

Claude 3.5 Sonnet (Anthropic, October 2024 release) was used as the primary intervention in this study to generate formative assessment questions from video lecture transcripts. The specific model and prompting strategies are described in detail in the Methods section and Multimedia Appendix 1. During manuscript preparation, the authors used ChatGPT (OpenAI, GPT-5.5, accessed June 2026) and Perplexity AI (accessed June 2026) to assist with language editing and to improve the clarity, readability, and English expression of the manuscript. These tools were not used to generate study data, perform analyses, determine study findings, create figures or tables, or generate references without independent verification. All AI-assisted suggestions were reviewed, edited, and verified by the authors, who take full responsibility for the content of the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The patient immunomonitoring course was developed and delivered as part of the regular educational activities of the Institute of Tropical Medicine (ITM), Antwerp. The ITM, Antwerp, had no influence on study design, data collection, analysis, or the decision to publish.

Authors' Contributions

Conceptualization: CKT, ZM, JR, JP-F, HJMN, D Katsuva, JMK, GL-N, CTS, SH, D Kannan, HH, LFG
Data curation: OG
Formal analysis: CKT
Methodology: CKT
Resources: CKT
Supervision: CKT, MZ
Validation: ED, OG, ARH-T, KF-M
Writing – original draft: CKT, ED, OG, ARH-T, KF-M, BKJ, MZ
Writing – review & editing: CKT, ED, OG, ARH-T, KF-M, BKJ, ZM, JR, JP-F, HJMN, D Katsuva, JMK, GL-N, CTS, SH, D Kannan, HH, LFG, MZ

Conflicts of Interest

None declared.

Multimedia Appendix 1

Prompt framework for AI-assisted question generation.
[DOCX File (Microsoft Word File), 22 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Student survey instrument.
[DOCX File (Microsoft Word File), 28 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Processing efficiency and review time supplementary data.
[DOCX File (Microsoft Word File), 28 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Examples of AI-generated assessment questions with expert review annotations.
[DOCX File (Microsoft Word File), 17 KB-Multimedia Appendix 4]

References

1. Roll I, Wylie R. Evolution and revolution in artificial intelligence in education. Int J Artif Intell Educ. 2016;26(2):582-599. [doi: 10.1007/s40593-016-0110-3]
2. El Khoury E. Mapping the response to AI and its impact on assessment redesign through document analysis. The assessment review. 2024. URL: https://assessatcuny.commons.gc.cuny.edu/2024/03/mapping-the-response-to-ai-and-its-impact-on-assessment-redesign-through-document-analysis/ [Accessed 2026-06-17]
3. Bennett RE. Educational assessment: what to watch in a rapidly changing world. Educ Meas Issues Pract. Dec 2018;37(4):7-15. [doi: 10.1111/emip.12231]
4. Luckin R, Holmes W, Griffiths M, Forcier LB. Intelligence unleashed: an argument for AI in education. Pearson; 2016. URL: https://static.googleusercontent.com/media/edu.google.com/en//pdfs/Intelligence-Unleashed-Publication.pdf [Accessed 2026-06-17]
5. Holmes W, Bialik M, Fadel C. Artificial Intelligence in Education Promises and Implications for Teaching and Learning. 1st ed. Center for Curriculum Redesign; 2019. ISBN: 9781794293700
6. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ. Jun 1, 2023;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
7. Coşkun Ö, Kıyak YS, Budakoğlu İİ. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: a randomized controlled experiment. Med Teach. Feb 2025;47(2):268-274. [doi: 10.1080/0142159X.2024.2327477] [Medline: 38478902]
8. Black P, Wiliam D. Assessment and classroom learning. Assess Educ Princ Policy Pract. 1998;5(1):7-74. [doi: 10.1080/0969595980050102]
9. Nicol DJ, Macfarlane-Dick D. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. Stud High Educ. 2006;31(2):199-218. [doi: 10.1080/03075070600572090]
10. Broadbent J, Poon WL. Self-regulated learning strategies and academic achievement in online higher education learning environments: a systematic review. Internet High Educ. Oct 2015;27:1-13. [doi: 10.1016/j.iheduc.2015.04.007]
11. Wisniewski B, Zierer K, Hattie J. The power of feedback revisited: a meta-analysis of educational feedback research. Front Psychol. 2019;10:3087. [doi: 10.3389/fpsyg.2019.03087] [Medline: 32038429]

12. Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. BMC Med Educ. Mar 29, 2024;24(1):354. [doi: 10.1186/s12909-024-05239-y] [Medline: 38553693]
13. Law AK, So J, Lui CT, et al. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. BMC Med Educ. Feb 8, 2025;25(1):208. [doi: 10.1186/s12909-025-06796-6] [Medline: 39923067]
14. Riehm L, Nanji K, Lakhani M, Pankiv E, Hasanee D, Pfeifer W. The use of large language models in generating multiple choice questions for health professions education: a systematic review and network meta-analysis. PLoS One. 2026;21(1):e0340277. [doi: 10.1371/journal.pone.0340277] [Medline: 41481658]
15. Wiggins GP, McTighe J. Understanding by Design. ASCD; 2005. ISBN: 9781416600350
16. Biggs J. Enhancing teaching through constructive alignment. High Educ. Oct 1996;32(3):347-364. [doi: 10.1007/BF00138871]
17. Mayer RE, Fiorella L. 19 - Principles for managing essential processing in multimedia learning: segmenting, pre-training, and modality principles. In: The Cambridge Handbook of Multimedia Learning. Cambridge University Press; 2021:243-260. [doi: 10.1017/9781108894333.025]
18. Paas F, van Merriënboer JJG. Cognitive-load theory: methods to manage working memory load in the learning of complex tasks. Curr Dir Psychol Sci. 2020;29(4):394-398. [doi: 10.1177/0963721420922183]
19. Liu D. The effects of segmentation on cognitive load, vocabulary learning and retention, and reading comprehension in a multimedia learning environment. BMC Psychol. Jan 2, 2024;12:4. [doi: 10.1186/s40359-023-01489-5] [Medline: 38167380]
20. Rey GD, Beege M, Nebel S, Wirzberger M, Schmitt TH, Schneider S. A meta-analysis of the segmenting effect. Educ Psychol Rev. 2019;31:389-419. [doi: 10.1007/s10648-018-9456-4]
21. Kiyan Tsunami C, Henríquez-Trujillo AR, Ferreira-Meyers K, et al. Guidelines for integrating actionable A-SMART learning outcomes into the backward design process. MedEdPublish (2016). 2024;14:242. [doi: 10.12688/mep.20606.1] [Medline: 39600516]
22. Sweller J. Cognitive load during problem solving: effects on learning. Cogn Sci. Apr 1988;12(2):257-285. [doi: 10.1207/s15516709cog1202_4]
23. Clark RC, Mayer RE. E-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning. 4th ed. Wiley; 2016. ISBN: 9781119158660
24. White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. In: Yoder J, Gabriel R, Vranić V, Brown K, editors. PLoP '23: Proceedings of the 30th Conference on Pattern Languages of Programs. 2023:1-31. [doi: 10.5555/3721041.3721046]
25. Koo M, Yang SW. Questionnaire use and development in health research. Encyclopedia. 2025;5(2):65. [doi: 10.3390/encyclopedia5020065]
26. Phillips AW, Friedman BT, Durning SJ. How to calculate a survey response rate: best practices. Acad Med. Feb 2017;92(2):269. [doi: 10.1097/ACM.0000000000001410] [Medline: 27655051]
27. Guo PJ, Kim J, Rubin R. How video production affects student engagement: an empirical study of MOOC videos. In: Sahami M, Fox A, Hearst MA, Chi MTH, editors. L@S '14: Proceedings of the First ACM Conference on Learning @ Scale Conference. Association for Computing Machinery; 2014:41-50. [doi: 10.1145/2556325.2566239]
28. Brame CJ. Effective educational videos: principles and guidelines for maximizing student learning from video content. CBE Life Sci Educ. 2016;15(4):es6. [doi: 10.1187/cbe.16-03-0125] [Medline: 27789532]
29. Tong L, Wang J, Rapaka S, Garg PS. Can ChatGPT generate practice question explanations for medical students, a new faculty teaching tool? Med Teach. Mar 2025;47(3):560-564. [doi: 10.1080/0142159X.2024.2363486] [Medline: 38900675]
30. Elzayyat M, Mohammad JN, Zaqout S. Assessing LLM-generated vs expert-created clinical anatomy MCQs: a student perception-based comparative study in medical education. Med Educ Online. Dec 2025;30(1):2554678. [doi: 10.1080/10872981.2025.2554678] [Medline: 40884796]
31. Zhang Q, Huang Z, Huang Y, et al. Generative AI in medical education: feasibility and educational value of LLM-generated clinical cases with MCQs. BMC Med Educ. Oct 27, 2025;25(1):1502. [doi: 10.1186/s12909-025-08085-8] [Medline: 41146115]
32. Arif T, Asthana S, Collins-Thompson K. Generation and assessment of multiple-choice questions from video transcripts using large language models. L@S '24 Proc ACM Conf Learn Scale. 2024:530-534. [doi: 10.1145/3657604.3664714]
33. Camarata T, McCoy L, Rosenberg R, Temprine Grellinger KR, Brettschnieder K, Berman J. LLM-generated multiple choice practice quizzes for preclinical medical students. Adv Physiol Educ. Sep 1, 2025;49(3):758-763. [doi: 10.1152/advan.00106.2024] [Medline: 40516963]

Abbreviations

A-SMART: action-oriented, specific, measurable, achievable, relevant, and time-based
ITM: Institute of Tropical Medicine
LLM: large language model
LO: learning outcome
MCQ: multiple-choice question
MR: multiple response
PEI: processing efficiency index
PIM: patient immunomonitoring
SME: subject-matter expert

Edited by Alicia Stone; peer-reviewed by Brenton T Bicknell, Inthrani Indran, Mayank Agarwal; submitted 24.Oct.2025; initial comments to author 13.May.2026; final revised version received 22.Sep.2026; accepted 23.Sep.2026; published 25.Nov.2026

Please cite as:
Kiyan Tsunami C, Diro E, Goovaerts O, Henríquez-Trujillo AR, Ferreira-Meyers K, Jacobs BKM, Mwanda Z, Rimal J, Pozu-Franco J, Montalvo Navarrete HJ, Katsuva Sibongwere D, Kolie JM, Luakanda-Ndelemo G, Semevo TC, Heng S, Kannan D, Hn H, Fucay Guin L, Zolfo M
Feasibility and Acceptability of AI-Generated Formative Assessments for Patient Immunomonitoring Training: Mixed Methods Implementation Study
JMIR Form Res 2026;10:e86309
URL: https://formative.jmir.org/2026/1/e86309
doi: 10.2196/86309

© Carlos Kiyan Tsunami, Ermias Diro, Odin Goovaerts, Aquiles Rodrigo Henríquez-Trujillo, Karen Ferreira-Meyers, Bart K M Jacobs, Ziyanda Mwanda, Jyotsna Rimal, Jamine Pozu-Franco, Héctor Javier Montalvo Navarrete, Deogratias Katsuva Sibongwere, Jean Michel Kolie, Gradi Luakanda-Ndelemo, Tchenagnon Claude Semevo, Sotheara Heng, Diljith Kannan, Harish Hn, Luis Fucay Guin, Maria Zolfo. Originally published in JMIR Formative Research (https://formative.jmir.org), 25.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.