

# Teaching Clinical Reasoning in the Age of AI: A Mixed-Methods Formative Evaluation of AI-Generated Script Concordance Tests and Expert Embodiment

Alexandre Hudon, Véronique Phan, Bernard Charlin, René Wittmer

Submitted to: JMIR Medical Education  
on: April 27, 2025

**Disclaimer:** © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 22

    Figures ..... 23

        Figure 1..... 24

    Multimedia Appendixes ..... 25

        Multimedia Appendix 1..... 26

    TOC/Feature image for homepages ..... 27

        TOC/Feature image for homepage 0..... 28

# Teaching Clinical Reasoning in the Age of AI: A Mixed-Methods Formative Evaluation of AI-Generated Script Concordance Tests and Expert Embodiment

Alexandre Hudon<sup>1, 2, 3, 4, 5</sup> BEng, MD, MSc(ed), PhD; Véronique Phan<sup>6, 7</sup> MD, MSc; Bernard Charlin<sup>5, 8</sup> MD, MEd, PhD; René Wittmer<sup>9</sup> MD

<sup>1</sup>Department of Psychiatry and Addictology Faculty of Medicine Université de Montréal Montréal CA

<sup>2</sup>Department of Psychiatry Institut universitaire en santé mentale de Montréal Montreal CA

<sup>3</sup>Centre de recherche de l'Institut universitaire en santé mentale de Montréal Montreal CA

<sup>4</sup>Department of Psychiatry Institut national de psychiatrie légale Philippe-Pinel Montreal CA

<sup>5</sup>Groupe Interdisciplinaire de recherche sur la cognition et le raisonnement professionnel (GIRCoPRo) Université de Montréal Montreal CA

<sup>6</sup>Department of Pediatrics Faculty of Medicine Université de Montréal Montreal CA

<sup>7</sup>Department of Pediatrics Centre Hospitalier Universitaire Sainte-Justine Montreal CA

<sup>8</sup>Centre de pédagogie appliquée aux sciences de la santé Faculty of Medicine Université de Montréal Montreal CA

<sup>9</sup>Department of family medicine and emergency medicine Faculty of Medicine Université de Montréal Montreal CA

## Corresponding Author:

Alexandre Hudon BEng, MD, MSc(ed), PhD

Department of Psychiatry and Addictology

Faculty of Medicine

Université de Montréal

Pavillon Roger-Gaudry, 2900 Bd Édouard-Montpetit Local L-315

Montréal

CA

## Abstract

**Background:** The integration of artificial intelligence (AI) in medical education is evolving, offering new tools to enhance teaching and assessment. Among these, script concordance tests (SCT) are well suited to evaluate clinical reasoning in contexts of uncertainty. Traditionally, SCTs require expert panels for scoring and feedback, which can be resource intensive. Recent advances in generative AI, particularly large language models (LLM), suggest the possibility of replacing human experts with simulated ones, though this potential remains underexplored.

**Objective:** This study aimed to evaluate whether LLMs can effectively simulate expert judgment in SCTs, by using generative AI to author, score, and provide feedback for SCTs in cardiology and pneumology. A secondary goal was to assess students' perceptions of the test's difficulty and the pedagogical value of AI-generated feedback.

**Methods:** A cross-sectional, mixed-methods study was conducted with 25 second-year medical students who completed a 32-item SCT authored by ChatGPT-4o. Six LLMs (three trained on course material and three untrained) served as simulated experts to generate scoring keys and feedback. Students answered SCT questions, rated perceived difficulty, and selected the most helpful feedback explanation for each item. Quantitative analysis included scoring, difficulty ratings, and correlation between student and AI responses. Qualitative comments were thematically analyzed.

**Results:** The average student score was 22.8 out of 32 (SD = 1.6), with scores ranging from 19.75 to 26.75. Trained AI systems showed significantly higher concordance with student responses ( $\kappa = 0.64$ ) than untrained models ( $\kappa = 0.41$ ). AI-generated feedback was rated as most helpful in 62.5% of cases, especially when provided by trained models. The SCT demonstrated good internal consistency (Cronbach's  $\alpha = 0.76$ ), and students reported moderate perceived difficulty (mean=3.7/7). Qualitative feedback highlighted appreciation for SCTs as reflective tools, while recommending clearer guidance on Likert-scale use and more contextual detail in vignettes.

**Conclusions:** This is among the first studies to demonstrate that trained generative AI models can reliably simulate expert clinical reasoning in a script concordance framework. The findings suggest that AI can both streamline SCT design and offer educational valuable feedback without compromising authenticity. Future studies should explore longitudinal effects on learning

and assess how hybrid models (human and AI) can optimize reasoning instruction in medical education.

(JMIR Preprints 27/04/2025:76618)

DOI: <https://doi.org/10.2196/preprints.76618>

## Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://preprints.jmir.org/preprint/76618>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://preprints.jmir.org/preprint/76618>

## Original Manuscript

## Original Paper

# Teaching Clinical Reasoning in the Age of AI: A Mixed-Methods Formative Evaluation of AI-Generated Script Concordance Tests and Expert Embodiment

## Abstract

### Background:

The integration of artificial intelligence (AI) in medical education is evolving, offering new tools to enhance teaching and assessment. Among these, script concordance tests (SCT) are well suited to evaluate clinical reasoning in contexts of uncertainty. Traditionally, SCTs require expert panels for scoring and feedback, which can be resource intensive. Recent advances in generative AI, particularly large language models (LLM), suggest the possibility of replacing human experts with simulated ones, though this potential remains underexplored.

### Objective:

This study aimed to evaluate whether LLMs can effectively simulate expert judgment in SCTs, by using generative AI to author, score, and provide feedback for SCTs in cardiology and pneumology. A secondary goal was to assess students' perceptions of the test's difficulty and the pedagogical value of AI-generated feedback.

### Methods:

A cross-sectional, mixed-methods study was conducted with 25 second-year medical students who completed a 32-item SCT authored by ChatGPT-4o. Six LLMs (three trained on course material and three untrained) served as simulated experts to generate scoring keys and feedback. Students answered SCT questions, rated perceived difficulty, and selected the most helpful feedback explanation for each item. Quantitative analysis included scoring, difficulty ratings, and correlation between student and AI responses. Qualitative comments were thematically analyzed.

### Results:

The average student score was 22.8 out of 32 (SD = 1.6), with scores ranging from 19.75 to 26.75. Trained AI systems showed significantly higher concordance with student responses ( $\rho = 0.64$ ) than untrained models ( $\rho = 0.41$ ). AI-generated feedback was rated as most helpful in 62.5% of cases, especially when provided by trained models. The SCT demonstrated good internal consistency (Cronbach's  $\alpha = 0.76$ ), and students reported moderate perceived difficulty (mean=3.7/7). Qualitative feedback highlighted appreciation for SCTs as reflective tools, while recommending clearer guidance on Likert-scale use and more contextual detail in vignettes.

### Conclusions:

This is among the first studies to demonstrate that trained generative AI models can reliably simulate expert clinical reasoning in a script concordance framework. The findings suggest that AI can both streamline SCT design and offer educational valuable feedback without compromising authenticity. Future studies should explore longitudinal effects on learning and assess how hybrid models (human and AI) can optimize reasoning instruction in medical education.

**Keywords:** Script concordance test; clinical reasoning; artificial intelligence; large language models; medical education; formative assessment; generative AI; expert simulation

## Introduction

One of the biggest challenges in medical education is evaluating clinical capabilities, especially higher-order abilities such as clinical reasoning and decision-making. Knowledge-based tests such as multiple-choice questions (MCQs) frequently fall short in capturing the subtleties of clinical reasoning in the face of diagnostic uncertainty [1]. By simulating actual clinical interactions, tools like workplace-based assessments and Objective Structured Clinical Examinations (OSCEs) try to assess clinical reasoning, but they come with a high logistical and human cost. An additional method for addressing thinking in clinical environments with unclear definitions is the script concordance test (SCT). In contrast to standard assessments, SCTs ask students to use a Likert-type scale to rate the impact of scenarios that are followed by additional clinical information on a postulated hypothesis. Concordance with a panel of expert responses serves as the basis for scoring, which reflects the variety and probabilistic character of actual clinical practice [2,3]. SCTs are valuable, but creating and scoring them requires time and involves hiring several specialists who must give careful, consistent answers to a variety of questions.

The format of a SCT usually involves presenting a clinical vignette followed by a diagnostic, investigative, or therapeutic hypothesis. A new piece of information is then introduced, and the test-taker must judge, using a Likert-type scale, how this new data affects the likelihood or relevance of the hypothesis. Responses are scored by comparing the learner's choices to those of a panel of experienced clinicians, with partial credit awarded based on the distribution of expert answers rather than a single correct response. This scoring system captures the variability in expert judgment and allows the SCT to measure concordance with expert reasoning rather than factual recall. Validity evidence for the SCT includes its ability to discriminate between levels of expertise, its positive impact on cognitive engagement, and its reliability when a sufficient number of items and panel members are used [3-5]. As such, SCTs are now used internationally in both undergraduate and postgraduate medical education, particularly in specialties that demand complex clinical judgment.

Artificial intelligence (AI) algorithms such as large language models (LLM) developments in recent years presented a possible paradigm change in this field. Having been trained on extensive text data sets, LLMs are deep learning models that can produce language which is similar to that of a person, summarize intricate concepts, and even mimic domain-specific reasoning [6]. These models, as opposed to rule-based systems, generate context-aware responses, often with good fluency, by identifying patterns in unstructured data. LLMs have demonstrated the ability to replicate patient or clinician discourse, produce high-quality feedback, and perform at or close to the passing threshold on USMLE-style questions [7-9]. By To create test questions, summarize clinical cases, and support self-directed learning, educators have started looking into LLMs.

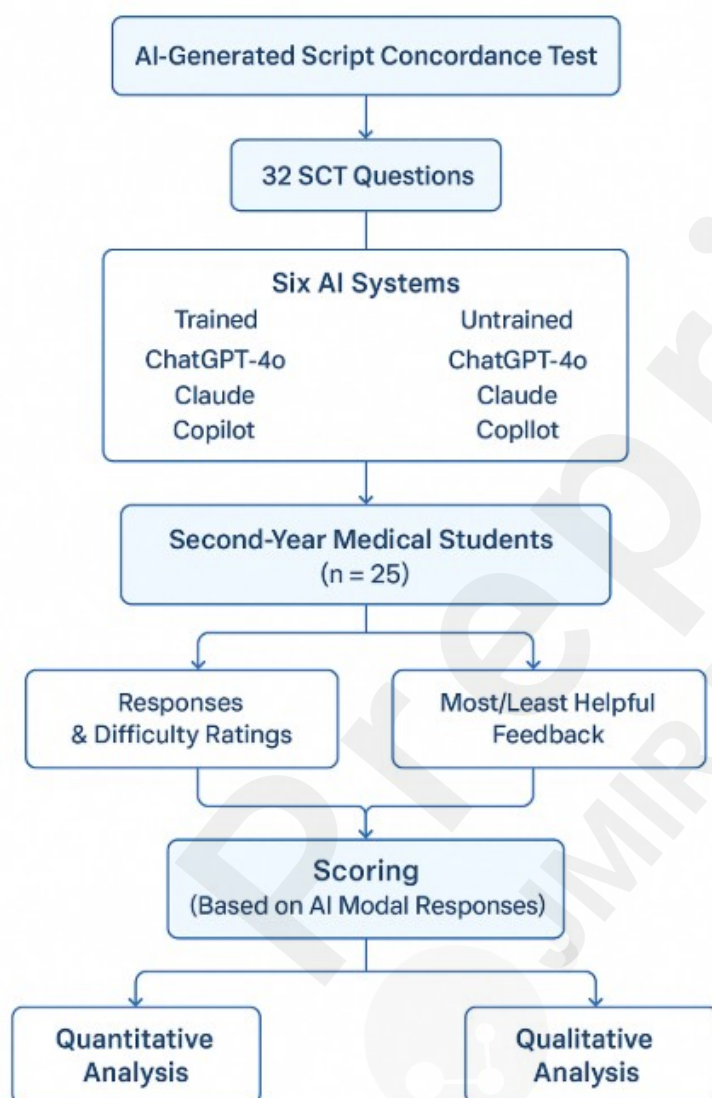
In the field of LLMs used to generate SCTs, several studies have demonstrated the usefulness of this tool for creating clinical vignettes in medical education and other health sciences fields [10,11]. A recent study highlighted the educational quality of script concordance vignettes created using ChatGPT (OpenAI) [11]. To use tools based on generative artificial intelligence in the design of educational materials, and to improve efficiency while optimizing human and material resources, it is essential to evaluate the level of difficulty of the proposed scenarios and questions [12]. Moreover, to our knowledge, no study has yet demonstrated the ability of generative artificial intelligence to embody the role of a clinical expert in selecting the most appropriate response to a given scenario and in providing feedback to learners.

The main objective of this project is to explore the performance of generative artificial intelligence in its ability to create script concordance tests for undergraduate medical students and to embody the role of an expert. A secondary objective is to evaluate how medical learners perceive the difficulty level of the AI-generated script concordance tests, as well as their level of appreciation for each type

of feedback provided.

## Methods

This study employed a cross-sectional, mixed-methods experimental design to evaluate the performance of medical students on a generative AI–authored formative SCT and to assess the use of AI systems as content experts for both scoring and providing formative feedback. This study was approved by the Université de Montréal’s Research Ethics Committee in Education and Psychology under project name 2024-6168. The flow of this study is presented in Figure 1.



**Figure 1.** Protocol Flow Diagram for this study.

## Participants and recruitment

The study was conducted in November 2024 at the Université de Montréal’s Faculty of Medicine. To be included in the study, participants needed to be enrolled in the second year of the undergraduate medical education program, to have completed their cardiology course, and to be preparing for their

respirology exam, having completed the related course materials. This timing allowed for the inclusion of clinical reasoning questions involving symptoms and signs that may relate to both systems, given their close physiological and clinical connections. An email was sent to all potential participants to register for this formative examination. Participation was anonymous and voluntary, with implied consent obtained through survey completion.

## Test development and structure

The script concordance test was constructed using generative AI (ChatGPT-4o) following best practices in SCT development [13]. The test comprised four clinical vignettes, each related to a different domain within the cardiology and pneumology curriculum: (1) chronic pulmonary obstructive disease (COPD) exacerbation, (2) hemoptysis, (3) chest pain, and (4) acute cough. Each vignette contained eight SCT items, for a total of 32 questions.

Each item followed the classic SCT structure:

- A clinical scenario introducing uncertainty;
- A diagnostic, investigative, or therapeutic hypothesis;
- A new clinical finding;

A 5-point Likert-type response scale indicating how the new information would influence the hypothesis (ranging from not likely to very likely).

Items were designed and reviewed to reflect varying degrees of diagnostic ambiguity, probabilistic reasoning, and real-world uncertainty.

The full test, in French and its English translation, is available in Multimedia Appendix 1. The following prompts were used to design the SCTs:

- Act as an expert in health sciences education at the university level, specializing in cardiology and pulmonology.
- Also act as an expert in designing script concordance test vignettes.
- Generate a script concordance vignette that includes eight questions based on a theme integrating both pulmonology and cardiology.
- Create questions related to the vignette that begin with: *"If you are considering 'a diagnostic hypothesis' and you find 'a sign or symptom', how does this finding affect the likelihood of the hypothesis?"* (using a Likert scale from -2 to +2 which refers to not likely to very likely).

## Use of Generative AI as Expert Panel

Rather than relying on human experts, this study used six LLMs as simulated expert panels. These models were selected as they were amongst the most popular at the time this study took place. These models were divided into two groups:

- Trained AI Experts (N=3): These systems (ChatGPT-4o, Claude 3.7, and Microsoft 365 Copilot) were provided with course-specific content, including lecture notes and clinical guidelines from the cardiopulmonary block, prior to test simulation.
- Untrained AI Experts (N=3): These were the same LLMs used without additional curricular material or fine-tuning, representing general-purpose versions of the models.

Each AI system was prompted to respond to all 32 SCT items as if it was a senior clinical expert. Their responses were used to establish scoring keys and to generate formative feedback explanations for each item. The modal response across all six AI experts served as the reference answer for scoring.

The prompts used to generate feedback were the same across the different AI systems. They were:

1. Act as a medical expert in cardiology and pneumology;
2. Act as a script concordance test expert;
3. For this vignette and each of its related questions ([Insert Vignette] and [Insert Questions]); please pick the most appropriate answer from the Likert-Scale;
4. For each of your answers, please provide an explanation as if you were trying to explain your reasoning to a medical student

## Student Completion and Data Collection

Students completed the SCT using an online Google Form, which included:

- The 32 SCT questions;
- A 7-point Likert scale for perceived difficulty for each question (from very easy to very hard);
- For each item, selection of the most helpful and least helpful feedback among six AI-generated rationales (labeled anonymously as “Expert 1” to “Expert 6”);
- Answering if they thought the question was AI-generated or not.
- Open-ended comment boxes following each question and at the end of the test.

The total estimated completion time was 35–45 minutes. Responses were exported and cleaned in Microsoft Excel for analysis.

## Scoring and Analysis

Student answers were scored using the aggregate partial credit method, whereby:

- 1 point was awarded if the student’s response matched the modal expert answer;
- Partial credit (e.g., 0.75, 0.5) was granted if the response matched a minority AI expert choice;
- 0 points were given if the response did not match any AI panel member.

Descriptive statistics were computed for total scores, individual item performance, and perceived difficulty. Internal consistency was measured using Cronbach’s alpha. AI-to-AI and AI-to-student concordance was evaluated using Spearman’s rank correlation  $\rho$ .

In the feedback evaluation component, student preferences for expert rationales were tabulated across all items. Each expert (AI system) received a score reflecting how frequently its feedback was selected as the “most helpful” and “least helpful.” This allowed comparison between trained and untrained AI systems as feedback providers.

## Qualitative Analysis

Open-ended responses were analyzed using an inductive thematic approach. Comments were reviewed, coded iteratively, and grouped into major and minor themes by two independent reviewers. Discrepancies were resolved through discussion and consensus. Themes focused on student perceptions of the SCT format, feedback quality, and suggestions for improvement.

## Results

### Participants

A total of 25 students volunteered to complete the SCT in a formative, non-evaluative setting. Each student also evaluated the difficulty of each question, selected the most and least helpful expert feedback, and provided qualitative comments on individual questions and on the overall examination. Six generative AI systems (three trained on course content, three untrained) completed the test and provided justifications for each item which served as immediate feedback for the students.

### Student performance

Among the 25 second-year medical students who completed the SCT, the average score was 22.8 out of 32 (standard deviation: 1.6), with individual scores ranging from 19.75 to 26.75. The distribution of scores approximated a normal curve, with a slight left skew, indicating generally strong alignment between student responses and the AI expert panel, while still highlighting variability in performance. This range suggests that the AI-generated SCT was capable of differentiating clinical reasoning abilities among participants, supporting its discriminatory power even in a formative context. The full distribution of the results for each question is found in Table 1.

Table 1. Distribution of the test results (N=25).

| Question (English)                                                                                     | mean | std  | min  | 25 % | 50 % | 75 % | max  |
|--------------------------------------------------------------------------------------------------------|------|------|------|------|------|------|------|
| Q1: To assess an acute COPD exacerbation, determine the impact of an oxygen saturation of 85%.         | 0,72 | 0,15 | 0,25 | 0,75 | 0,75 | 0,75 | 1    |
| Q2: To evaluate acute pulmonary edema, judge the impact of a recent respiratory infection.             | 0,65 | 0,2  | 0    | 0,75 | 0,75 | 0,75 | 0,75 |
| Q3: To assess the hypothesis of pulmonary embolism, determine the impact of dyspnea and tachycardia.   | 0,76 | 0,17 | 0,25 | 0,75 | 0,75 | 0,75 | 1    |
| Q4: To evaluate myocardial dysfunction, judge the influence of a high heart rate.                      | 0,56 | 0,27 | 0    | 0,5  | 0,75 | 0,75 | 1    |
| Q5: To differentiate pulmonary edema from asthma exacerbation, assess the importance of a chest X-ray. | 0,89 | 0,21 | 0,5  | 1    | 1    | 1    | 1    |

|                                                                                                                              |      |          |          |          |          |          |          |
|------------------------------------------------------------------------------------------------------------------------------|------|----------|----------|----------|----------|----------|----------|
| Q6: To evaluate a COPD exacerbation, determine how a history of smoking modifies the hypothesis.                             | 0,82 | 0,1<br>5 | 0,5      | 0,7<br>5 | 0,7<br>5 | 1        | 1        |
| Q7: To guide diagnosis in respiratory distress, assess the importance of history-taking.                                     | 0,97 | 0,1<br>1 | 0,5      | 1        | 1        | 1        | 1        |
| Q8: To diagnose pulmonary edema, judge the influence of acute EKG ischemia.                                                  | 0,79 | 0,1<br>9 | 0,2<br>5 | 0,7<br>5 | 0,7<br>5 | 1        | 1        |
| Q9: To assess tuberculosis as a differential diagnosis, evaluate the impact of night sweats and weight loss.                 | 0,78 | 0,0<br>8 | 0,7<br>5 | 0,7<br>5 | 0,7<br>5 | 0,7<br>5 | 1        |
| Q10: To assess the likelihood of chronic pulmonary hemoptysis, judge the influence of a history of bronchiectasis.           | 0,76 | 0,1<br>3 | 0,2<br>5 | 0,7<br>5 | 0,7<br>5 | 0,7<br>5 | 1        |
| Q11: To assess life-threatening hemoptysis requiring urgent care, evaluate the impact of an oxygen saturation of 92%.        | 0,59 | 0,3<br>2 | 0        | 0,2<br>5 | 0,7<br>5 | 0,7<br>5 | 1        |
| Q12: To evaluate bronchopulmonary cancer, determine the influence of a 30 pack-year smoking history.                         | 0,86 | 0,1<br>3 | 0,7<br>5 | 0,7<br>5 | 0,7<br>5 | 1        | 1        |
| Q13: To orient toward an infectious etiology like tuberculosis or pneumonia, assess the usefulness of the patient's history. | 1    | 0        | 1        | 1        | 1        | 1        | 1        |
| Q14: To distinguish possible causes of hemoptysis, assess the importance of performing a chest X-ray.                        | 1    | 0        | 1        | 1        | 1        | 1        | 1        |
| Q15: To assess active bronchiectasis, judge the interaction between smoking history and risk of complications.               | 0,82 | 0,1<br>4 | 0,5      | 0,7<br>5 | 0,7<br>5 | 1        | 1        |
| Q16: To diagnose tuberculosis-related hemoptysis, evaluate the influence of prior tuberculosis exposure.                     | 0,86 | 0,1<br>3 | 0,7<br>5 | 0,7<br>5 | 0,7<br>5 | 1        | 1        |
| Q17: To assess acute coronary syndrome (ACS), evaluate the impact of ST elevation in DII, DIII, and aVF.                     | 0,99 | 0,0<br>5 | 0,7<br>5 | 1        | 1        | 1        | 1        |
| Q18: To evaluate atypical ACS presentation, assess the influence of diabetes.                                                | 0,75 | 0,1<br>6 | 0,5      | 0,7<br>5 | 0,7<br>5 | 0,7<br>5 | 1        |
| Q19: To assess aortic dissection as a cause of pain, evaluate the impact of hypertension.                                    | 0,77 | 0,1      | 0,5      | 0,7<br>5 | 0,7<br>5 | 0,7<br>5 | 1        |
| Q20: To evaluate suspected pneumothorax, judge how pain radiation influences the hypothesis.                                 | 0,29 | 0,2<br>7 | 0        | 0,2<br>5 | 0,2<br>5 | 0,2<br>5 | 1        |
| Q21: To differentiate between ACS and pulmonary embolism, assess the importance of the EKG.                                  | 0,96 | 0,1<br>2 | 0,5      | 1        | 1        | 1        | 1        |
| Q22: To evaluate pulmonary embolism, determine the influence of elevated blood pressure on the hypothesis.                   | 0,55 | 0,2<br>3 | 0,2<br>5 | 0,2<br>5 | 0,7<br>5 | 0,7<br>5 | 0,7<br>5 |
| Q23: To assess ACS likelihood, judge how chest pain radiation differs between male and                                       | 0,77 | 0,2      | 0,2      | 0,7      | 0,7      | 1        | 1        |

|                                                                                                                                  |      |          |          |          |          |          |     |
|----------------------------------------------------------------------------------------------------------------------------------|------|----------|----------|----------|----------|----------|-----|
| female patients.                                                                                                                 |      |          | 5        | 5        | 5        |          |     |
| Q24: To determine the urgency of management, assess the influence of diabetes and hypertension.                                  | 0,78 | 0,2<br>1 | 0,5      | 0,5      | 0,7<br>5 | 1        | 1   |
| Q25: To assess community-acquired pneumonia, evaluate the impact of initial fever on the hypothesis.                             | 0,86 | 0,1<br>3 | 0,7<br>5 | 0,7<br>5 | 0,7<br>5 | 1        | 1   |
| Q26: To evaluate treatment-induced cough, determine the influence of ACE inhibitor use.                                          | 0,72 | 0,1<br>8 | 0,2<br>5 | 0,7<br>5 | 0,7<br>5 | 0,7<br>5 | 1   |
| Q27: To assess severe pneumonia requiring hospitalization, evaluate the impact of a normal oxygen saturation (97%).              | 0,15 | 0,1<br>6 | 0        | 0        | 0,2<br>5 | 0,2<br>5 | 0,5 |
| Q28: To assess acute bronchitis, evaluate the impact of crackles on the hypothesis.                                              | 0,43 | 0,2<br>8 | 0        | 0,2<br>5 | 0,2<br>5 | 0,7<br>5 | 1   |
| Q29: To distinguish between ACE-inhibitor-induced cough and acute respiratory infection, assess the usefulness of the history.   | 1    | 0        | 1        | 1        | 1        | 1        | 1   |
| Q30: To assess upper respiratory tract infection, evaluate the influence of a 10-day duration of cough.                          | 0,7  | 0,2      | 0,2<br>5 | 0,7<br>5 | 0,7<br>5 | 0,7<br>5 | 1   |
| Q31: To differentiate pneumonia from bronchitis, assess the importance of a chest X-ray.                                         | 1    | 0        | 1        | 1        | 1        | 1        | 1   |
| Q32: To guide the need for antibiotic treatment in suspected pneumonia, evaluate the role of initial fever and productive cough. | 0,84 | 0,2      | 0,2<br>5 | 0,7<br>5 | 1        | 1        | 1   |

A more detailed item-level analysis revealed that questions with clearer clinical cues such as Q1, which evaluated the impact of hypoxemia (SpO<sub>2</sub> of 85%) on the hypothesis of COPD exacerbation, and Q8, which examined the influence of EKG findings on the diagnosis of acute pulmonary edema. These questions had high rates of agreement with the expert panel. In these items, over 80% of student responses matched the most common expert answer, suggesting that students were comfortable applying well-established diagnostic patterns. In contrast, questions such as Q2 (regarding the influence of a recent respiratory infection on the hypothesis of pulmonary edema) and Q4 (assessing the relevance of tachycardia for myocardial dysfunction) showed greater variability in student responses, with less than 50% aligning with the modal expert choice. These items appeared to introduce more clinical ambiguity or demanded a higher level of inference, which may explain the broader dispersion of answers. Preliminary psychometric analysis demonstrated a Cronbach's alpha of 0.76, indicating good internal consistency across the 32 items.

About 40% of students scored 25 or above, frequently aligning with expert responses even on more ambiguous questions. These students tended to report lower perceived difficulty and used the "slightly confirmed" response category more judiciously. The remaining 60% scored between 19 and 24 points and were more likely to select "non influenced" or "slightly infirmed" options. This may reflect a more cautious reasoning style or uncertainty in applying new clinical information. Overall, the SCT demonstrated its utility in capturing variations in reasoning patterns and levels of diagnostic confidence among early clinical learners.

## Agreement Between AI and the Medical Students

To evaluate the extent to which LLM systems could simulate expert reasoning in SCTs, the answers of the six AI systems were compared with those of the medical students. The results demonstrated a difference in the alignment between trained and untrained AI systems with medical students' responses. Among the trained models, ChatGPT-4o achieved the highest agreement, with a correlation coefficient  $\rho$  of 0.68 ( $p < 0.001$ ) compared to the medical students. Claude and Microsoft Copilot followed closely, with correlations of 0.64 and 0.61, respectively (both  $p = 0.017$ ). These values suggest a high level of concordance in clinical reasoning between trained AI and medical students, particularly in questions anchored in familiar pathophysiological concepts (such as the diagnostic value of ST-segment elevation in myocardial infarction or oxygen saturation in COPD exacerbation). On average, trained AI systems yielded a correlation coefficient of 0.64, reflecting a strong and consistent tendency to emulate clinical judgment when appropriately contextualized with curricular information.

In contrast, the untrained AI models demonstrated lower levels of agreement. The untrained version of ChatGPT-4o showed a moderate correlation ( $\rho = 0.48$ ,  $p = 0.041$ ), while Claude and Copilot, when untrained, showed weaker and statistically non-significant correlations of 0.42 ( $p = 0.062$ ) and 0.34 ( $p = 0.081$ ), respectively. These models tended to exhibit less consistent patterns, often defaulting to neutral or non-committal answers such as "non influenced," especially in items requiring nuanced interpretation of clinical signs. Their reasoning lacked the contextual anchoring present in the trained models and showed more variability across similar clinical scenarios. On average, the untrained models produced a correlation coefficient of 0.41, substantially lower than their trained counterparts.

## Perceived difficulty

In addition to completing the 32 SCT items, students were asked to rate the perceived difficulty of each question on a 7-point Likert scale, ranging from 1 ("very easy") to 7 ("very difficult"). This measure provided insight into how learners experienced the complexity and cognitive demands of the AI-generated test items. The mean perceived difficulty across all questions was 3.7, suggesting that students generally found the test to be of moderate difficulty. This aligns with the formative purpose of the SCT and indicates that the AI-generated items were accessible yet still challenging enough to stimulate clinical reasoning.

When analyzed at the item level, some questions emerged as consistently more difficult. In particular, Q2 (which assessed how a recent respiratory infection influences the hypothesis of acute pulmonary edema) and Q4 (which explored the relevance of tachycardia in the context of suspected myocardial dysfunction) received higher average difficulty ratings, with mean scores exceeding 4.5. These questions involved ambiguous or indirect relationships between the new information and the hypothesis, requiring students to reason within clinical gray zones where data interpretation is not straightforward. In contrast, items that presented clear diagnostic anchors, such as Q1 (hypoxemia in COPD) and Q8 (EKG ischemia in pulmonary edema), were rated as significantly easier, with average difficulty scores below 3.0. These findings suggest that item clarity and familiarity with pathophysiological mechanisms strongly influence perceived complexity.

Furthermore, there was no statistically significant difference in perceived difficulty between items identified by students as being written by artificial intelligence versus those believed to be authored

by humans (t-test,  $p = 0.47$ ). This finding supports the face validity of AI-generated items, indicating that they are perceived as authentic and comparable in challenge to traditional, faculty-written SCT items. Several students even remarked in open comments that they were "unable to tell" which questions had been generated by AI, further supporting the indistinguishability of AI-authored content when properly crafted.

Cluster analysis of difficulty ratings also revealed subtle differences in student subgroups. High-performing students (as defined by SCT scores  $\geq 25/32$ ) tended to rate the test as less difficult overall (mean=3.3) compared to their peers (mean=3.9), which may reflect their greater familiarity with clinical reasoning or more efficient script activation. Additionally, free-text comments frequently noted that clinical ambiguity (rather than item length or language) was the primary factor contributing to perceived difficulty. Students consistently emphasized that the challenge came from "figuring out what matters most" in the presence of partial or conflicting clinical cues. The difficulty ratings offer an important validation metric for AI-generated SCTs. The overall moderate difficulty, the range of item-specific variation, and the lack of perceived discrepancy between AI- and human-generated content all point toward a well-calibrated test.

## Evaluation of AI-expert feedback

In addition to rating the difficulty of each SCT item, students were asked to select the expert feedback they considered most helpful and least helpful among the options provided. For every SCT question, six expert rationales were made available to the medical students by the AI systems (trained and untrained). Students were instructed to base their selection on three specific criteria: clarity, clinical relevance, and overall educational value. This evaluation provided insight not only into the perceived credibility of AI-generated explanations, but also into the comparative strengths of different expert types in formative feedback contexts.

Across all 32 questions, AI-generated feedback (especially from the models trained on the course material) was selected as helpful in 62.5% of instances. Among these, ChatGPT-4o (trained) was most frequently chosen, followed by Claude (trained) and Copilot (trained). Students frequently cited the clarity of language, conciseness, and alignment with taught clinical reasoning strategies as reasons for preferring these explanations. Trained AI feedback was especially appreciated in items requiring pathophysiological reasoning, such as those related to acute coronary syndromes and differential diagnoses in respiratory presentations. The consistent use of structured logic and evidence-based phrasing contributed to a perception of trustworthiness and educational quality.

In contrast, the feedback generated by untrained AI systems was only selected as most helpful in 9.4% of instances, and more frequently identified as the least useful. The untrained models often produced generic or excessively cautious explanations, relying on vague phrases such as "may support the hypothesis" or "needs further investigation" without directly engaging with the specific clinical cues provided in the vignette. These models also occasionally failed to recognize important pathophysiological links, leading to misalignment with students' expectations. Several students noted that while the untrained AI responses were not necessarily incorrect, they lacked the pedagogical precision needed for effective feedback.

## Qualitative Feedback from Students

To complement the quantitative analyses, students were invited to provide open-ended comments after each SCT question and at the end of the test. These narrative responses were analyzed using an inductive thematic analysis. A total of 94 discrete comments were submitted by the 25 participating students. From these, three major themes were identified: (1) ambiguity and interpretation of Likert-scale options, (2) the educational value of SCTs, and (3) recommendations for improvement in item design and feedback delivery.

### ***Ambiguity and Interpretation of Likert-Scale Options***

A recurrent theme in the comments related to students' uncertainty in differentiating between closely related answer choices. Many expressed difficulties in deciding whether a new piece of clinical information "slightly confirmed" versus "confirmed" a hypothesis, particularly when the vignette lacked strong diagnostic anchors.

- "It's really hard to know when to say 'confirmed' versus 'slightly confirmed' as they feel almost the same without lab results or imaging."
- "Sometimes the answer depends on how I imagine the rest of the case. Without more detail, it's hard to be confident."

This theme reflects a broader cognitive tension between probabilistic reasoning and binary decision-making in early clinical training. Students seemed to grasp the script concordance philosophy but struggled with the semantic granularity required by the scale, particularly in low-certainty situations.

### ***Educational Value of the SCT Format***

Despite difficulties with Likert scales, students generally expressed strong appreciation for the SCT format as a learning tool. Many found it stimulated reflection on clinical uncertainty and offered a welcome departure from the right/wrong dichotomy typical of multiple-choice assessments.

- "This test made me think like a real clinician. It's not just about memorizing, it's about understanding probability."
- "I liked how the questions made me weigh the value of each piece of information, not just jump to a diagnosis."

Students also praised the opportunity to compare their reasoning to experts, particularly when feedback was provided in structured, explanatory formats. Some noted that this type of test encouraged them to revisit and revise their clinical scripts, reinforcing clinical reasoning pathways rather than isolated facts.

### ***Recommendations for Improvement***

Several students offered constructive suggestions aimed at improving the test format and clarity. One prominent comment was the request for more context in the vignettes, such as vital signs, lab values, or imaging findings. Students felt that additional details would allow them to apply reasoning with greater precision and confidence.

- "More case details would help me justify my choice better. Sometimes it felt like guessing."

- “It would be helpful to have the patient’s age or comorbidities as it changes how I interpret the clue.”

Another frequently mentioned point was the desire for guidance or training on how to use the Likert scale. Some students suggested including sample questions or expert rationales before the actual test to calibrate their expectations.

- “Before the real test, I would have liked an example with expert explanation as it would help me feel less lost.”

Interestingly, several students highlighted their surprise at the quality of feedback provided by AI systems. These comments were often framed with curiosity or mild skepticism, indicating an openness to AI as a pedagogical tool, contingent on quality and relevance.

- “I thought one of the best explanations came from AI. I couldn’t even tell at first.”
- “The AI answers were clear and to the point. Honestly better than some professors I’ve had.”

While students appreciated the reflective and authentic nature of the format, they also identified areas where clearer structure or training could enhance their experience. Interestingly, their willingness to accept AI-generated feedback, when well-crafted, suggests that such technologies could play a productive role in clinical education, particularly when embedded within thoughtful educational frameworks. This thematic analysis suggests future iterations of the SCT should include:

- A calibration phase with annotated examples;
- More contextual information in vignettes;
- Continued refinement of AI-generated feedback to preserve clarity, conciseness, and clinical congruence.

## Discussion

### Principal Results

This study explores the performance of generative artificial intelligence in its ability to create script concordance tests for undergraduate medical students and to embody the role of an expert models by fully designing, scoring, and providing feedback for a SCT in undergraduate medical education. Among the 25 second-year medical students who completed the AI-generated SCT, the mean score was 22.8 out of 32, with a distribution that reflected meaningful variation in clinical reasoning patterns. Students generally aligned well with expert-modeled responses, particularly on items with strong pathophysiological anchors, while questions requiring greater inferential reasoning yielded more dispersed answers. Trained AI systems (those provided with course-specific materials) achieved the highest concordance with both student performance and expected reasoning pathways, offering feedback that was frequently rated as the most helpful. Untrained AI models, in contrast, were perceived as less pedagogically effective and exhibited weaker correlations with student responses. The SCT also performed reliably, with a Cronbach’s alpha of 0.76, and was rated by students as moderately difficult, further reinforcing the test’s capacity to capture meaningful distinctions in clinical reasoning.

## Comparison with Prior Work

The integration of AI in formative clinical assessment tools such as SCTs is still at an early stage, but this study provides one of the first real-world demonstrations that AI systems can meaningfully simulate expert reasoning when adequately trained. While prior studies have explored LLMs' ability to answer MCQs or USMLE-style questions, those formats tend to rely on recall and fixed answers [14,15]. In contrast, SCTs demand probabilistic thinking and tolerance of ambiguity, qualities that better reflect real-world medical clinical reasoning and thus could be used in exams to test clinical reasoning. The ability of trained LLMs in this study to mirror the distribution of expert responses (and to provide feedback that students found clear and pedagogically relevant) suggests a step forward in the use of AI not merely as a content generator, but as a cognitive proxy for experienced clinicians.

This study also aligns with findings from recent work examining AI's capacity to replicate reflective thinking patterns. For instance, Nori et al. highlighted that LLMs like GPT-4 are increasingly capable of "meta-cognitive" behavior, adapting their justifications based on context and complexity [16]. The findings support this, especially in items where the trained AI systems offered nuanced rationales that resonated with learners. Furthermore, the indistinguishability between human- and AI-authored feedback, as perceived by students in this study, aligns with previous reports by Lee et al., who found that students often could not discern between AI-generated and expert-written content when both were of high quality [17]. Importantly, students in this cohort demonstrated openness to AI-generated feedback, provided it was contextually accurate and clearly structured, suggesting that AI can be accepted as a trusted educational voice when properly curated.

Nonetheless, this project also illustrates the current limitations of general-purpose LLMs. Untrained models, though capable of generating grammatically correct responses, frequently defaulted to cautious or generic reasoning, especially in questions requiring interpretation of subtle clinical cues. This reduced their pedagogical value and undermined student trust. These findings are consistent with recent critiques from Arora et al., who argue that untrained LLMs lack the contextual grounding needed for expert-level interpretation, particularly in nuanced or culturally specific clinical scenarios [18]. As such, effective integration of AI into medical education likely hinges on customized fine-tuning, prompt engineering, and ongoing human oversight to ensure that AI explanations remain accurate, relevant, and educationally meaningful.

## Limitations

This study has several limitations. The sample size was modest and drawn from a single institution, which may limit generalizability. The test was also formative in nature, and students may not have engaged with the same level of cognitive effort as they would in summative settings. It is also possible that solely students who expressed an interest in AI accepted to participate to this SCT which may affect the generalizability of the qualitative components of this study. Additionally, while the AI models were trained on course materials, their responses still depended on prompt clarity and token limits, which may have shaped their output in subtle ways. No direct comparison was made between AI and human expert panels on the same test, limiting conclusions about relative superiority. Finally, as this was a single-round implementation, long-term effects on clinical reasoning or retention were not assessed. Future studies should explore how repeated exposure to AI-generated SCTs impacts learning trajectories, and whether hybrid models (AI and human feedback) offer added value over AI alone.

## Conclusions

As artificial intelligence technologies evolve and become increasingly embedded in medical education, this study offers novel insight into how large language models can assume the role traditionally occupied by human experts in clinical reasoning assessments. By designing and scoring an SCT entirely with generative AI systems, and by leveraging both trained and untrained models to simulate expert feedback, we demonstrated that AI can produce assessment content that is not only functionally valid but also well-received by learners. The high degree of alignment between student responses and trained AI-generated reference panels, alongside the positive reception of AI-authored feedback, suggests that these technologies hold real promise for supporting formative assessment. While distinctions remain between the effectiveness of trained versus untrained models, this study illustrates that with proper calibration, AI systems can enhance both the efficiency and pedagogical depth of clinical education. These findings open the door for further research into scalable, AI-assisted assessment models that can flexibly support reasoning under uncertainty which is an essential competency in modern medical practice.

## Conflicts of Interest

None declared

## Abbreviations

ACS: acute coronary syndrome  
AI: artificial intelligence  
COPD: chronic obstructive pulmonary disease  
EKG: electrocardiogram  
JMIR: Journal of Medical Internet Research  
LLM: large language model  
MCQ: multiple-choice question  
RCT: randomized controlled trial  
SCT: script concordance test  
USMLE: United States Medical Licensing Examination

## Data Availability Statements

The individual test results analyzed during this study are available from the corresponding author on reasonable request.

## Multimedia Appendix

Multimedia Appendix 1 : Script concordance test (French and English).

## References

1. Schuwirth LW, Van der Vleuten CP. Programmatic assessment: From assessment of learning to assessment for learning. *Med Teach*. 2011;33(6):478-85. doi: 10.3109/0142159X.2011.565828. PMID: 21609177.
2. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The Script Concordance test: a tool to assess the reflective clinician. *Teach Learn Med*. 2000 Fall;12(4):189-95. doi:

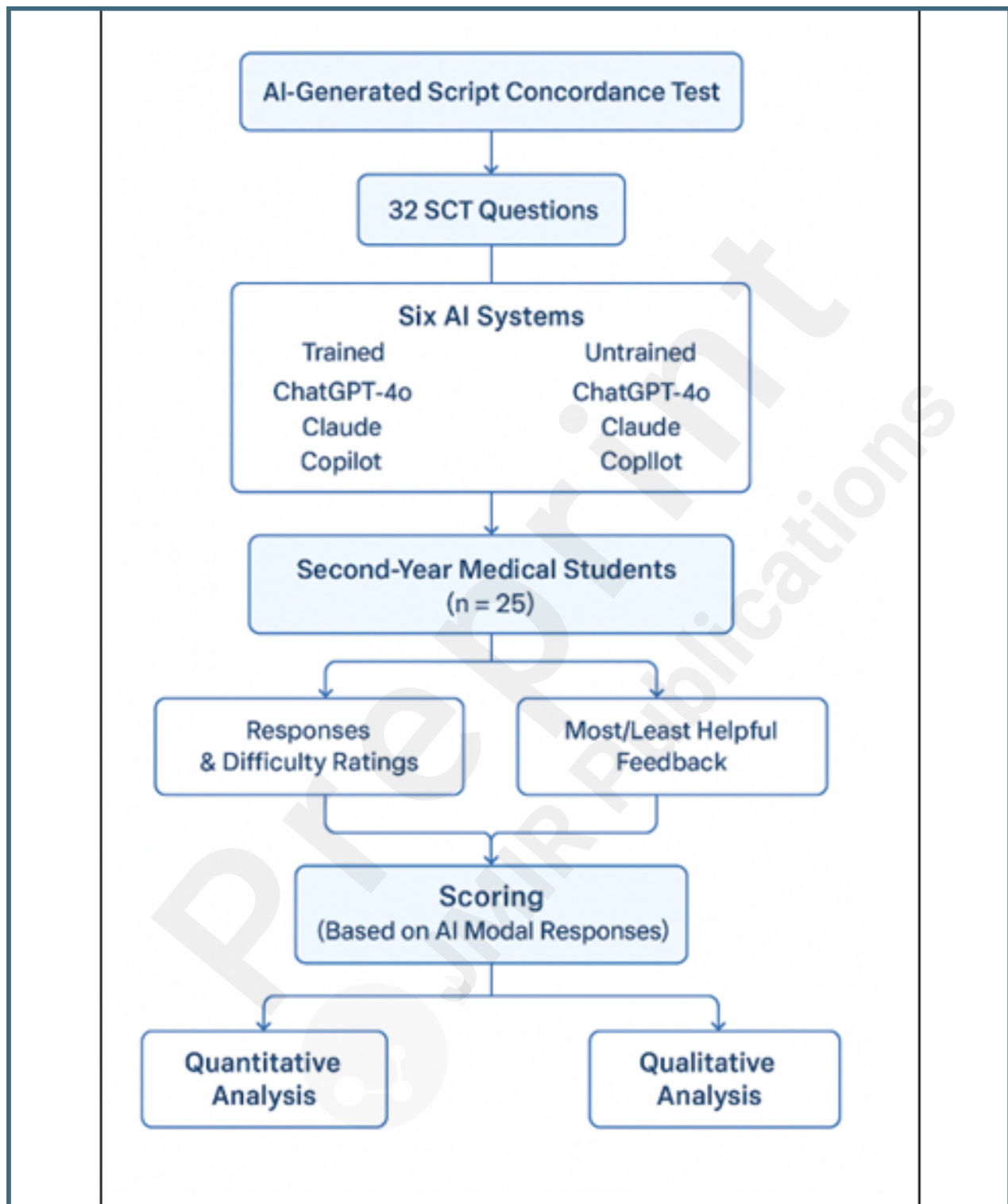
- 10.1207/S15328015TLM1204\_5. PMID: 11273368.
3. Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CP. Script concordance testing: a review of published validity evidence. *Med Educ*. 2011 Apr;45(4):329-38. doi: 10.1111/j.1365-2923.2010.03863.x. PMID: 21401680.
  4. Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: insights from a systematic review. *Med Educ*. 2012 Jun;46(6):552-63. doi: 10.1111/j.1365-2923.2011.04211.x. PMID: 22626047.
  5. Gagnon R, Charlin B, Coletti M, Sauvé E, van der Vleuten C. Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? *Med Educ*. 2005 Mar;39(3):284-91. doi: 10.1111/j.1365-2929.2005.02092.x. PMID: 15733164.
  6. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Agüera Y Arcas B, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A, Natarajan V. Large language models encode clinical knowledge. *Nature*. 2023 Aug;620(7972):172-180. doi: 10.1038/s41586-023-06291-2. Epub 2023 Jul 12. Erratum in: *Nature*. 2023 Aug;620(7973):E19. doi: 10.1038/s41586-023-06455-0. PMID: 37438534; PMCID: PMC10396962.
  7. Jeblick K, Schachtner B, Dextl J, Mittermeier A, Stüber AT, Topalis J, Weber T, Wesp P, Sabel BO, Ricke J, Ingrisich M. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol*. 2024 May;34(5):2817-2825. doi: 10.1007/s00330-023-10213-1. Epub 2023 Oct 5. PMID: 37794249; PMCID: PMC11126432.
  8. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, Aziz S, Damseh R, Alabed Alrazak S, Sheikh J. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. *JMIR Med Educ*. 2023 Jun 1;9:e48291. doi: 10.2196/48291. PMID: 37261894; PMCID: PMC10273039.
  9. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023 Feb 9;2(2):e0000198. doi: 10.1371/journal.pdig.0000198. PMID: 36812645; PMCID: PMC9931230.
  10. Yanagita Y, Yokokawa D, Fukuzawa F, Uchida S, Uehara T, Ikusaka M. Expert assessment of ChatGPT's ability to generate illness scripts: an evaluative study. *BMC Med Educ*. 2024 May 15;24(1):536. doi: 10.1186/s12909-024-05534-8. PMID: 38750546; PMCID: PMC11095028.
  11. Hudon A, Kiepora B, Pelletier M, Phan V. Using ChatGPT in Psychiatry to Design Script Concordance Tests in Undergraduate Medical Education: Mixed Methods Study. *JMIR Med Educ*. 2024 Apr 4;10:e54067. doi: 10.2196/54067. PMID: 38596832; PMCID: PMC11007379.
  12. Choi GW, Kim SH, Lee D, Moon J. Utilizing Generative AI for Instructional Design: Exploring Strengths, Weaknesses, Opportunities, and Threats. *TechTrends*. 2024;68(3):832-844. doi:10.1007/s11528-024-00967-w.
  13. Fournier JP, Demeester A, Charlin B. Script concordance tests: guidelines for construction. *BMC Med Inform Decis Mak*. 2008 May 6;8:18. doi: 10.1186/1472-6947-8-18. PMID: 18460199; PMCID: PMC2427021.
  14. Luo D, Liu M, Yu R, et al. Evaluating the performance of GPT-3.5, GPT-4, and GPT-4o in the Chinese National Medical Licensing Examination. *Sci Rep*. 2025;15:14119. doi:10.1038/s41598-025-98949-2.
  15. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How Does

- ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ.* 2023 Feb 8;9:e45312. doi: 10.2196/45312.
16. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv [Preprint]*. 2023 Apr 12:arXiv:2303.13375. doi:10.48550/arXiv.2303.13375.
  17. Lee S, Song KS. Teachers' and students' perceptions of AI-generated concept explanations: Implications for integrating generative AI in computer science education. *Comput Educ Artif Intell.* 2024;7:100283. doi:10.1016/j.caeai.2024.100283.
  18. Arora A, Arora A. The promise of large language models in health care. *Lancet.* 2023 Feb 25;401(10377):641. doi: 10.1016/S0140-6736(23)00216-7. PMID: 36841609.

## Supplementary Files

## Figures

Protocol Flow Diagram for this study.



## **Multimedia Appendixes**

Script concordance test (French and English).

URL: <http://asset.jmir.pub/assets/8654f8f8f1145549e010f92eab28d74a.docx>



## **TOC/Feature image for homepages**

JMIR Placeholder.

[ PLACEHOLDER ]