Revista Española
de Educación Médica

UNIVERSIDAD DE MURCIA

Using Large Language Models to Generate Script Concordance Test in Medical Education: ChatGPT and Claude

Yavuz Selim Kıyak[1] *, **Emre Emekli**[2]

1 Department of Medical Education and Informatics, Gazi University Faculty of Medicine, Ankara, Turkiye; yskiyak@gazi.edu.tr, 0000-0002-5026-3234

2 Department of Radiology, Faculty of Medicine, Eskişehir Osmangazi University, Eskişehir, Turkiye; 0000-0001-5989-1897

* Correspondence: yskiyak@gazi.edu.tr

Received: 4/11/24; Accepted: 2/12/24; Published: 3/12/24

Abstract

We aimed to determine the quality of AI-generated (ChatGPT-4 and Claude 3) Script Concordance Test (SCT) items through an expert panel. We generated SCT items on abdominal radiology using a complex prompt in large language model (LLM) chatbots (ChatGPT-4 and Claude 3 (Sonnet) in April 2024) and evaluated the items' quality through an expert panel of 16 radiologists. Expert panel, which was blind to the origin of the items provided without modifications, independently answered each item and assessed them using 12 quality indicators. Data analysis included descriptive statistics, bar charts to compare responses against accepted forms, and a heatmap to show performance in terms of the quality indicators. SCT items generated by chatbots assess clinical reasoning rather than only factual recall (ChatGPT: 92.50%, Claude: 85.00%). The heatmap indicated that the items were generally acceptable, with most responses favorable across quality indicators (ChatGPT: 71.77%, Claude: 64.23%). The comparison of the bar charts with acceptable and unacceptable forms revealed that 73.33% and 53.33% of the questions in the items can be considered acceptable, respectively, for ChatGPT and Claude. The use of LLMs to generate SCT items can be helpful for medical educators by reducing the required time and effort. Although the prompt provides a good starting point, it remains crucial to review and revise AI-generated SCT items before educational use. The prompt and the custom GPT, "Script Concordance Test Generator", available at https://chatgpt.com/g/g-RlzW5xdc1-script-concordance-test-generator, can streamline SCT item development.

Keywords: script concordance test; clinical reasoning; medical education; artificial intelligence; chatgpt

1. Introduction

Clinical reasoning is an important skill in medical education. It is central for students and professionals to deal with the unpredictable nature of clinical practice. There are various methods to assess clinical reasoning skills (1). Among these methods, Script Concordance Test (SCT) has a particular emphasis on uncertainty (2-3) and therefore it provides significant advantages. However, creating SCT items is a challenging task.

Each SCT item begins with a brief clinical vignette that presents a patient's health condition. Following this, a key option about the scenario is presented. Then a piece of additional information is provided, such as, a new symptom, results of a test, and the effect of a treatment. The impact of this additional information on the initial option is assessed. Examinees provide their response by using a five-point Likert scale to determine whether the new finding is positive, negative, or neutral in terms of the appropriateness of the initial option, and to what extent (2–4). In scoring, these responses are compared with the responses of a panel of experts. SCTs are used for formative and summative assessment, applied across specialties, and aids in undergraduate and postgraduate training. They uniquely assess clinical reasoning under uncertainty by presenting learners with ambiguous clinical scenarios and asking them to evaluate the impact of new, often uncertain, information on the initial considerations, therefore resembling real-world clinical decision-making processes where information is often incomplete or evolving. This complex nature of an SCT item demands substantial effort from experts in the development process. Therefore, automatization of this process can bring efficiency.

AI in medical education has a wide range of applications (5). One area that has been explored in recent studies is the potential of AI in generating clinical cases (6–9) and multiple-choice questions (MCQs). A recent literature review revealed the promising results in using the large language model (LLM) chatbots for this purpose in various medical fields (10). Specifically in radiology, a study demonstrated that these chatbots are able to generate board-style MCQs (11). However, there is a gap when it comes to AI-driven generation of SCT items. To date, there is only one publication that has studied this area by focusing on the use of ChatGPT for SCT generation (12). The pioneering study for SCT generation with AI has some limitations. The prompt utilized in the study is lacking in detail to some extent, while good prompting is a crucial factor for more effective outputs. Additionally, its scope is narrowly tailored to psychiatric diagnostics, which restricts its broader application in diverse medical contexts.

To address these gaps, a group of researchers developed a complex prompt template (13) for generating SCT items that allows customization based on specific needs. In this study, we aimed to determine the quality of AI-generated (ChatGPT-4 and Claude 3) SCT items on abdominal radiology through an expert panel.

2. Methods

The study did not formally include any human participants and any interventions, therefore, it does not require us to apply for obtaining an ethical approval. We generated SCT items by using the detailed prompt template (13) in ChatGPT-4 and Claude 3 (Sonnet). Table 1 presents the prompt template, the specifications we added to the template, and an example of the generated SCT items.

We focused on the investigation of five types of pain related to abdominal radiology: Left lower quadrant pain (LLQP), right lower quadrant pain (RLQP), right upper quadrant pain (RUQP), epigastric pain (EP), and acute onset flank pain (AOFP). We generated an SCT item for each type of pain both in ChatGPT and Claude. We generated all items in April 2024 by establishing a new conversation page through their public websites for each item. The items are available in the Supplementary Material.

We created an expert panel in order to evaluate the quality of SCT items. This panel consisted of 16 radiologists because the ideal number of panel members is 15-20 (2). It included two professors, five specialists, and nine residents with at least three years of experience. The average medical experience of the panel members was 9.72 ± 7.42 years.

We provided 10 SCT items (each with three questions) to the panel without any modifications. Table 2 presents one of the SCTs. We did not add the answers and explanations provided by the LLMs. We also provided a brief information sheet about SCT method as a reminder. For each item, we provided 12 statements for evaluating the quality of items, together with three response options: yes, no, I am not sure. These statements were based on the "Script Concordance Test item grid" developed by Fournier (2) as were previously used in the study by Hudon et al. (12):

Table 1. The prompt template for generating script concordance test items using large language model chatbots, the specifications used for the study.

You are a script concordance test (SCT) developer for medical exams for [GENERAL DESCRIPTION OF THE TARGET GROUP, SUCH AS, "undergraduate medical students", "postgraduate medical trainees", "continuing medical education participants", OR A MORE DETAILED INFORMATION, SUCH AS "last year radiology residents in Turkey"].

SCT is a method used to assess clinical reasoning and decision-making skills in healthcare students and professionals. It is designed to evaluate how they interpret clinical information and make decisions under conditions of uncertainty. Each scenario is a clinical vignette describing a medical situation. For each scenario, there are multiple possible options or actions that a physician could take or consider in that situation.

Each SCT consists of four main components.

1. A very brief clinical vignette on a clinical problem regarding [TYPE WHAT THE SET OF QUESTIONS SHOULD FOCUS ON: E.G. diagnosis, order of investigation, treatment].

In a single table with four rows, the first row includes the labels:

2. First column: ONLY three different key plausible [TYPE WHAT THE SET OF QUESTIONS SHOULD FOCUS ON: E.G. diagnosis, order of investigation, treatment] for the clinical vignette. The label of the column should be "If you were thinking of …".

3. Second column: A new piece of clinical information for each of the initial option to make the situation significantly more complex based on the clinical vignette, such as, new symptoms, previously undisclosed aspects of the patient's history, physical examination findings, and results of tests or previous treatments. The label of the column should be "Then you learn that …".

4. Third column: Asking the examinee to rate on a 5-point scale (from -2 to +2) how this new information impacts or affects the initial option - whether it makes that option more or less appropriate, and to what degree. The label of the column should be "It becomes … (-2 much less appropriate, -1 less appropriate, 0 no change, +1 more appropriate, +2 much more appropriate)".

So the examinee has to reason through how the new clinical data changes or updates the appropriateness of the initial [TYPE WHAT THE SET OF QUESTIONS SHOULD FOCUS ON: E.G. diagnosis, order of investigation, treatment] option they were given. The 5-point scale allows them to indicate if the new information makes the original option more likely, less likely, or doesn't really change it, and to what extent. The idea is that this tests the student's clinical reasoning abilities by presenting evolving, uncertain medical cases and having them re-evaluate options as new information comes in, which is what happens in real medical practice.

Write a SCT for the following clinical problem: [TYPE A CLINICAL PROBLEM OR TOPIC OR LEARNING OBJECTIVE RELATED TO DIAGNOSIS, INVESTIGATION, OR TREATMENT] according to [TYPE THE NAME OF A RELEVANT GUIDELINE OR JUST TYPE "the reliable guidelines and/or expert consensus on this clinical problem"]. Indicate the level of appropriateness in the third column cells, together with a very brief rationale.

Completion of the prompt template with following details:
Target group: "undergraduate medical students"
Questions' focus: "order of investigation"
Guideline: "the reliable guidelines and/or expert consensus on this clinical problem"
Symptom (filled separately): "left lower quadrant pain", "right lower quadrant pain", "right upper quadrant pain", "epigastric pain", "acute onset flank pain"

Table 2. An example script concordance test item for "right upper quadrant pain" generated by ChatGPT-4.

A 45-year-old female presents with acute onset of right upper quadrant pain. The pain is sharp, worsens with movement, and is associated with nausea but no vomiting. There is no history of similar pain. The patient denies any recent travel or changes in medication.						
If you were thinking of ...	Then you learn that ...	It becomes ... (-2 much less appropriate, -1 less appropriate, 0 no change, +1 more appropriate, +2 much more appropriate)				
A. Ordering an abdominal ultrasound	The patient has a history of gallstones and the pain radiates to the right shoulder.	-2	-1	0	1	2
B. Ordering liver function tests (LFTs)	The patient's skin and eyes are jaundiced, and she mentions dark urine.	-2	-1	0	1	2
C. Ordering a CT scan of the abdomen	The patient has a slight fever and leukocytosis is noted on a CBC test.	-2	-1	0	1	2

a. The scenario described a challenging situation, even for experts.
b. The scenario was of appropriate difficulty for medical students.
c. Reading the scenario was necessary to understand the question and context.
d. The clinical presentation in the scenario was typical.
e. The scenario was appropriately written.
f. The questions (presented in sets of three options) were prepared to include important aspects related to the topic.
g. Options were relevant to the case.
h. The same option did not appear in two consecutive questions.
i. It was possible to test the relationship between the added new information (2nd column) and the initially presented option (1st column).
j. The Likert scale (3rd column) was clearly and explicitly defined.
k. The questions were designed to distribute answers evenly across all values of the Likert scale.
l. The questions allowed for balanced ambiguity, enabling different interpretations by experts.

In addition to those, we asked whether the items are for assessing only factual recall or clinical reasoning skill. The panel members, blind to the items' origins, independently provided their answers and evaluated each item using hard copies printed on paper.

We used RStudio and Python for data analysis and visualization. Along with simple descriptive statistics and bar charts for response distribution in terms of Likert scale (-2, -1, 0, 1, 2), we created a heatmap for each question based on the scenarios, responses (yes, no, I am not sure (uncertain)) to 12 statements, and the LLMs (ChatGPT, Claude). More "yes" and fewer "no" responses by experts to these statements point out that the item is of higher quality, although it is not realistic to expect "yes" from all experts for an item.

We also compared the bar charts of the response distributions with the acceptable and unacceptable forms provided by Lubarsky et al. (3), as presented in Figure 2 (charts in the middle):

- Acceptable1: Ideal variability of responses, higher discriminatory power.
- Acceptable2: Presence of a "deviant" response, removing such responses does not impact score reliability.

- Unacceptable1: Unanimity of responses; unacceptable due to resembling single best answer MCQs.
- Unacceptable2: Uniform divergence of responses; indicates non-discriminatory item.

3. Results

In the evaluation carried out by 16 radiologists, most of them stated that the generated five items assess clinical reasoning rather than factual recall, ChatGPT: 74/80 (92.50%), Claude: 68/80 (85.00%). We presented the results regarding 12 statements on the quality of items in Figure 1 as a heatmap. Although the findings varied item to item, it showed that the scenarios were not difficult for experts (statement "a"). The difficulty was somewhat appropriate for undergraduate medical students (statement "b"). For statements other than "a" and "b", the darkest cells mostly found in the "yes" area, and the white cells in the "no" area.

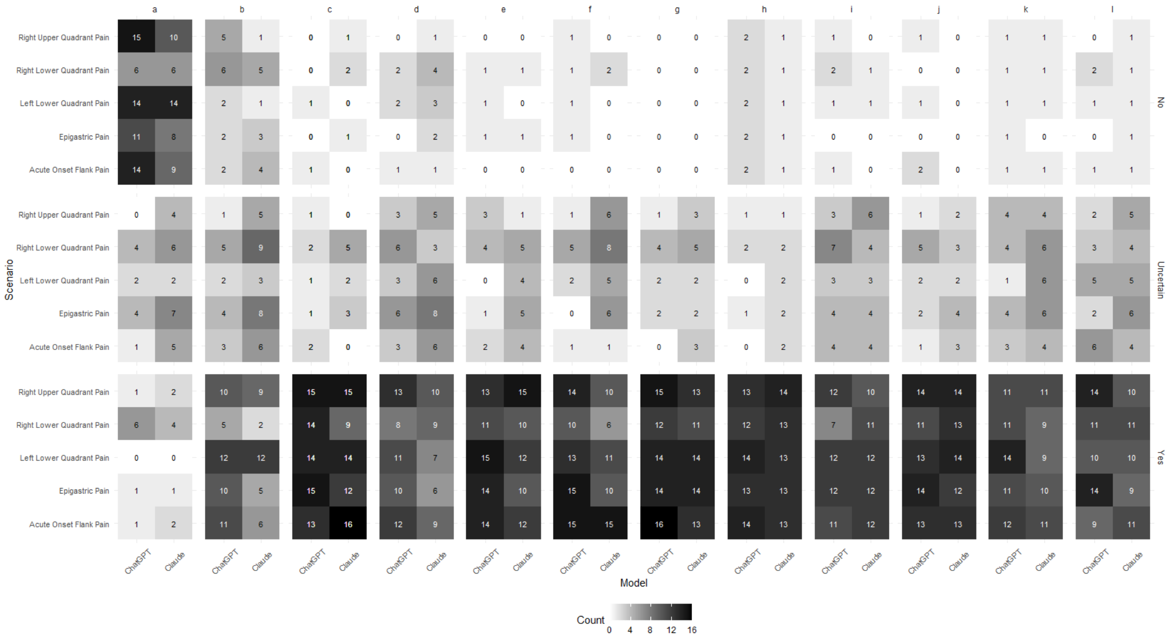

Figure 1. Heatmap that visualizes the distribution and frequency of responses from 16 experts on the evaluation of AI-generated, ChatGPT-4 and Claude 3 (Sonnet), script concordance test items across a range of medical scenarios and evaluative statements labeled from "a" to "l". Each cell in the heatmap represents the count of a specific type of response ("No", "I am not sure (Uncertain)", "Yes") to a statement. Those chosen more by the panel members are darker. A greater number of "Yes" responses from experts indicate higher item quality, although it is unrealistic to expect unanimous agreement from all experts.

In total, ChatGPT-generated SCT items received 12.40% (119) "no", 15.83% (152) "uncertain", and 71.77% (689) "yes" for 12 statements. In contrast, Claude 3 (Sonnet) received 10.01% (96) "no", 25.76% (247) "uncertain", and 64.23% (616) "yes". This indicates that ChatGPT-4 provided a higher proportion of acceptance overall.

Figure 2 includes the bar charts constructed based on responses in the expert panel to the questions. The comparison of these bar charts with acceptable and unacceptable forms, which have been described in the methods section, revealed that 11/15 (73.33%) and 8/15 (53.33%) of the questions in the five items can be considered acceptable for ChatGPT and Claude, respectively.

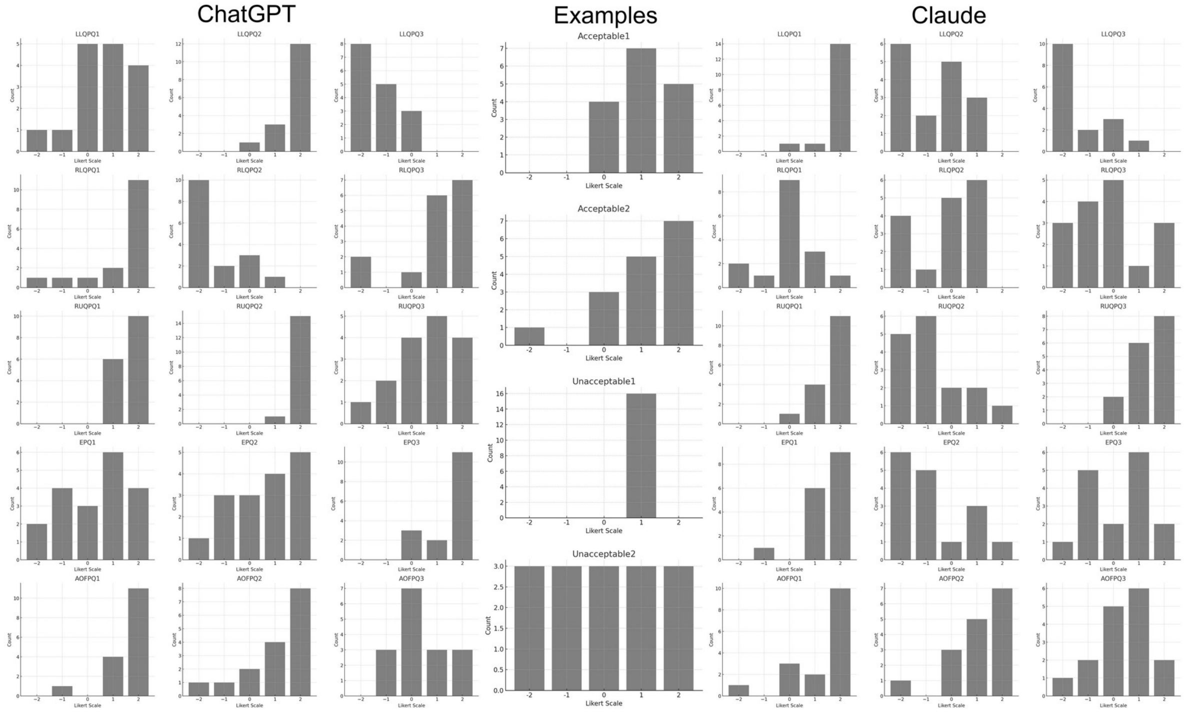

Figure 2. The series of bar charts displays the responses of 16 experts to five script concordance test items per AI model, ChatGPT-4 and Claude 3 (Sonnet). Each item consisted of three questions, with responses captured on a Likert scale ranging from -2 to 2. The charts are organized into three sections, one for each model (left: ChatGPT, right: Claude), and include a central comparison section featuring examples of acceptable and unacceptable forms adapted from Lubarsky et al. (2013). For example, ChatGPT's LLQPQ3 chart was similar to (mirroring) "Acceptable1" chart, and Claude's AOFPQ2 chart was similar to "Acceptable2" chart. However, as ChatGPT's RUQPQ2 was similar to "Unacceptable1" chart (item similar to single best answer MCQs), and ChatGPT's EPQ1 to "Unacceptable2" (non-discriminatory item)

4. Discussion

Building upon the previous study (12), the findings demonstrated that ChatGPT-4 and Claude 3 (Sonnet) models are able to generate acceptable SCT items when the detailed prompt is used. The findings align with the fact that ChatGPT-4 outperforms Claude 3 (Sonnet) in various benchmarks. However, newer models, such as Claude 3.5 (Sonnet) and OpenAI's o1 model, could provide different outcomes. In both of the chatbots, the response distribution was not ideal in each SCT. Our findings show that it is still important to review AI-generated SCT items before using them in educational settings, as the current capabilities of LLMs generates hallucinations (14) and struggle with managing tasks that require complex reasoning (15). When reviewing AI-generated SCT items, 12 statements can be a good rubric. For instance, both chatbots were very good at providing relevant options, as none of the evaluators chose "no" in that statement. However, many evaluators were uncertain or disagreed about the scenarios being appropriately challenging for experts and suitable in difficulty for students, which indicates that these aspects require revisions. Since generating SCTs using the prompt with different clinical problems and target groups may yield different strengths and limitations than we found, all 12 criteria should be used to review AI-generated SCT items. Despite the limitations, LLMs can still streamline the item writing process by providing a draft for experts to begin with.

Our study is limited to 10 SCT items on five abdominal radiology pain presentations, generated by ChatGPT and Claude LLM chatbots, and based only on expert evaluations in a single center without comparisons with human-written SCTs. Therefore, more studies are needed in different topics for more generalized inferences on the effectiveness of LLMs in generating SCT items. Using the prompt to generate more SCT items and administering them in various educational settings would allow investigation of evidence from different sources of validity.

Researchers can use the custom GPT (16), titled “Script Concordance Test Generator”, which is accessible through https://chatgpt.com/g/g-RlzW5xdc1-script-concordance-test-generator, to generate SCT items without the need to copy and paste the prompt each time. A cost analysis of LLM-assisted and human-written SCTs, similar to the one conducted by Lam et al. (9) for clinical cases, could be helpful for medical schools to understand its financial impact.

Moreover, although the prompt and custom GPT provide a good starting point, different prompting techniques (17) and workflows can improve the quality of outputs. For instance, agent-based systems can significantly improve output quality in LLMs (18). Therefore, an agent workflow could be developed to iteratively refine the items, instead of relying only on a prompt. These systems could also automate the analysis of student responses and provide personalized feedback. Integrating these systems with electronic SCT platforms would enable scalable, personalized assessment and feedback delivery that tailors educational content to each learner's needs. Additionally, while we used two proprietary models, open-source LLMs such as Llama, Mistral, and Command R+ deserve attention. These models allow for customization based on our specific needs and enable running them locally to mitigate privacy and security issues.

Educators should keep in mind that LLMs are progressing very quickly. As newer models become available, they need to incorporate and evaluate newer models. As a useful strategy, they could monitor the general performance using an online leaderboard (https://chat.lmsys.org/?leaderboard) based on over one million human pairwise comparisons of the models. Since each LLM have specific advantages and limitations, their capability in generating SCTs should be evaluated for each new model by benefiting from standardized evaluation metrics. Cross-model comparisons and error analysis are necessary before using them to ensure robust application of up-to-date models but critically spot idiosyncrasies and/or errors between each LLM.

5. Conclusions

- LLMs (ChatGPT-4 and Claude 3) demonstrated potential in generating acceptable SCT items for abdominal radiology in order to assess clinical reasoning and streamline item development in medical education.
- Although LLMs provide a helpful starting point, expert review and revision are essential to refine SCT items for appropriate difficulty and quality.
- Continued advancements in LLMs and enhanced workflows, including iterative prompting and agent-based refinement, could further optimize SCT generation, while newer models and open-source alternatives should be explored for broader applicability and customization.

Supplementary material: The prompt template to generate SCT items and the generated SCT items using ChatGPT-4 and Claude 3 (Sonnet) is available via this link: https://zenodo.org/records/14034424. **Figures**: Figures 1 and 2 are also available in high resolution.

Funding: There has been no funding.

Declaration of conflict of interest: The authors declare that they have no conflict of interest.

Author Contributions: YSK: Conceptualization, Methodology, Statistical Analysis, Data Visualization, Writing - original draft, Writing - review & editing; EE: Conceptualization, Data Collection, Writing - review & editing

Data Availability: The dataset can be accessed via Zenodo at https://zenodo.org/records/14034424.

References

1. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. Acad Med. 2019 Jun; 94(6):902–12. https://doi.org/10.1097/acm.0000000000002618
2. Fournier JP, Demeester A, Charlin B. Script Concordance Tests: Guidelines for Construction. BMC Med Inform Decis Mak. 2008 Dec;8(1):18. https://doi.org/10.1186/1472-6947-8-18

3. Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: From theory to practice: AMEE Guide No. 75. Med Teach. 2013 Mar;35(3):184–93. https://doi.org/10.3109/0142159x.2013.760036
4. Lubarsky S, Charlin B, Cook DA, Chalk C, Van Der Vleuten CPM. Script concordance testing: a review of published validity evidence: Validity evidence for script concordance tests. Med Educ. 2011 Apr;45(4):329–38. https://doi.org/10.1111/j.1365-2923.2010.03863.x
5. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach. 2024 Apr 2;46(4):446–70. https://doi.org/10.1080/0142159x.2024.2314198
6. Bakkum MJ, Hartjes MG, Piët JD, Donker EM, Likic R, Sanz E, et al. Using artificial intelligence to create diverse and inclusive medical case vignettes for education. Brit J Clinical Pharma. 2024 Jan 6;90(3):640–8. https://doi.org/10.1111/bcp.15977
7. Coşkun Ö, Kıyak YS, Budakoğlu Iİ. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: A randomized controlled experiment. Med Teach. 2024 Mar 13; https://doi.org/10.1080/0142159x.2024.2327477
8. Cook DA. Creating virtual patients using large language models: scalable, global, and low cost. Med Teach. 2024 Jul 11; https://doi.org/10.1080/0142159x.2024.2376879
9. Lam G, Shammoon Y, Coulson A, Lalloo F, Maini A, Amin A, et al. Utility of large language models for creating clinical assessment items. Med Teach. 2024 Aug 26;1–5. https://doi.org/10.1080/0142159x.2024.2382860
10. Kıyak YS, Emekli E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. Postgrad Med J. 2024 Jun 6; https://doi.org/10.1093/postmj/qgae065
11. Mistry NP, Saeed H, Rafique S, Le T, Obaid H, Adams SJ. Large Language Models as Tools to Generate Radiology Board-Style Multiple-Choice Questions. Acad Radiol. 2024 Jul;S107663322400432X. https://doi.org/10.1016/j.acra.2024.06.046
12. Hudon A, Kiepura 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–e54067. https://doi.org/10.2196/54067
13. Kıyak YS, Emekli E. A Prompt for Generating Script Concordance Test Using ChatGPT, Claude, and Llama Large Language Model Chatbots. Revista Española de Educación Médica. 2024;5(3):1–8. https://doi.org/10.6018/edumed.612381
14. Masters K. Medical Teacher's first ChatGPT's referencing hallucinations: Lessons for editors, reviewers, and teachers. Med Teach. 2023 Jul;45(7):673–5. https://doi.org/10.1080/0142159x.2023.2208731
15. Al-Naser Y, Halka F, Ng B, Mountford D, Sharma S, Niure K, et al. Evaluating Artificial Intelligence Competency in Education: Performance of ChatGPT-4 in the American Registry of Radiologic Technologists (ARRT) Radiography Certification Exam. Academic Radiology. 2024 Aug;S1076633224005725. https://doi.org/10.1016/j.acra.2024.08.009
16. Masters K, Benjamin J, Agrawal A, MacNeill H, Pillow MT, Mehta N. Twelve tips on creating and using custom GPTs to enhance health professions education. Med Teach. 2024 Jan 29;46(6):752–6. https://doi.org/10.1080/0142159x.2024.2305365
17. Kıyak YS. Beginner-Level Tips for Medical Educators: Guidance on Selection, Prompt Engineering, and the Use of Artificial Intelligence Chatbots. Med Sci Educ. 2024 Aug 17; https://doi.org/10.1007/s40670-024-02146-1
18. Li J, Wang S, Zhang M, Li W, Lai Y, Kang X, et al. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents [Internet]. arXiv; 2024 [cited 2024 May 10]. Available from: http://arxiv.org/abs/2405.029571