Revista Española de Educación Médica

UNIVERSIDAD DE MURCIA

A Prompt for Generating Script Concordance Test Using ChatGPT, Claude, and Llama Large Language Model Chatbots

Yavuz Selim Kıyak[1]*, Emre Emekli[2]

1 Department of Medical Education and Informatics, Gazi University Faculty of Medicine, Ankara, Turkey; yskiyak@gazi.edu.tr, , 0000-0002-5026-3234
2 Department of Radiology, Faculty of Medicine, Eskişehir Osmangazi University, Eskişehir, Turkey; 0000-0001-5989-1897
* Correspondence: yskiyak@gazi.edu.tr

Received: 16/4/24; Accepted: 15/5/24; Published: 16/5/24

Abstract

Medical education always evolves to incorporate more tools for specific needs in assessing clinical reasoning skills. Among these tools, Script Concordance Test (SCT) has a particular importance due to its focus on assessing decision-making in uncertain clinical situations. However, development of SCT items is effortful. Artificial intelligence tools, such as large language models, offer significant benefits. These models are already used for generating multiple-choice questions, and their use in generating SCTs offers great promise. However, this requires well-designed prompts to generate SCTs. This article proposes a generic prompt for the ChatGPT-4, Claude 3, Llama 3, and ChatGPT-4o large language model chatbots to generate SCTs, which can be tailored to various fields of medicine and different stages of medical education. It can help to streamline the development process of SCTs. Initial findings are promising, and there is a need for generating SCTs using large language models and conducting research to assess the quality of SCTs.

Keywords: script concordance test; clinical reasoning; medical education, artificial intelligence, ChatGPT, automatic item generation

Introduction

Clinical reasoning is the cognitive process by which clinicians gather information, diagnose, and manage patient care, starting from the initial encounter and continuing beyond treatment completion (1). It is one of the most important skills in medical education. There are various written assessment tools to assess clinical reasoning, such as, multiple-choice questions (MCQs), Extended Matching Questions (EMQs), Key Feature Questions (KFQs), and Script Concordance Test (SCT) (2). Among these tools, SCT is different because it pays a particular attention to uncertainty in assessing reasoning skills (3). Given that uncertainty is central to clinical practice and should be included in medical curriculum (4-5), SCT deserves more attention in medical education.

In an SCT item, examinees are given a series of medical scenarios, each forming a separate case with its own set of questions, designed to assess clinical reasoning skills under uncertain conditions (6). Each SCT item starts with a brief clinical vignette that summarizes a patient's condition or problem. Then, a plausible key option/action (e.g. diagnosis, investigation, treatment) related to the scenario is provided. Next, examinee is shown a new piece of information, such as, a symptom, a pre-existing condition, a result from an imaging study, or a laboratory test. Based on this new information, examinees need to evaluate how it affects the initial option/action. They use a five-point Likert scale to indicate whether the new finding has a positive, negative, or neutral impact on the appropriateness of that, and the level of this impact (7).

Although it has advantages such as assessing reasoning in uncertainty and focusing on the reasoning process over facts, developing high-quality SCT items is a complex and resource-

intensive process. It requires significant time and effort to construct these test items (8). Therefore, some medical teachers find it confusing and difficult to develop (9). As we are in the post-ChatGPT era (10), there is a potential to benefit from large language models (LLMs) in the development process of high-quality SCTs. However, good outputs from LLMs require the use of good prompts. Although there are well-designed prompts for generating MCQs using ChatGPT (11,12) and a growing body of evidence on the validity of MCQs generated by ChatGPT (13–17), there is a lack of prompts and research in terms of generating SCT items using LLMs. As of April 2024, to our best knowledge, there is only one study that focused on using an LLM chatbot (ChatGPT-3.5) for SCT generation (18). The pioneering study states that their prompt has some limitations (18) and therefore there is a need for more complex and well-designed prompt to generate better SCT items using LLMs. Moreover, the prompt has been designed to generate only diagnostic SCTs in psychiatry (18). Therefore, a more generic prompt is needed to allow users to tailor it to the needs of different fields and various purposes beyond diagnosis.

Due to this gap in medical education practice and literature, we aimed to present our prompt for generating SCT items with LLM chatbots in medical education. We demonstrated this using ChatGPT-4, Claude 3 (Opus), Llama 3 (Meta-Llama-3-70B-Instruct), and ChatGPT-4o.

The prompt and the generated SCT items

The prompt that we developed (table 1) is based on two key studies on developing SCT (6-7) and by considering the previous prompts (11,18). Since our prompt template is generic and can be tailored based on the needs of users, the user should fill the corresponding parts, which have been shown in bold characters between square brackets, with the following details:

1. **Target Group**: Are these SCT items for undergraduate medical students, postgraduate medical trainees, continuing medical education participants, or another specific group? The first part in bold characters should be filled with this information.
2. **Focus of the Questions**: As the literature recommends that each SCT should focus on a single aspect—be it diagnosis, investigation, or treatment (6), the user must specify the focus. Which aspect do you need about a clinical problem: diagnosis, order of investigation, or treatment? The following three parts in the prompt template, which is "[TYPE WHAT THE SET OF QUESTIONS SHOULD FOCUS ON: E.G. diagnosis, order of investigation, treatment])" highlighted in bold, should consistently be filled with your choice among these options. For example, if you want to generate an SCT on treatment, you should type "treatment" in each one of these parts. If you want to generate one more SCT item but on investigation, you should type "order of investigation" in each one of these three parts.
3. **Clinical Problem or Topic**: What is the clinical problem or topic the SCT should focus on? This part should be filled with a clinical problem/topic (e.g., Right Upper Quadrant Pain) or a learning objective.
4. **Guideline**: If there is a specific guideline that the user wants to base the questions on, it can be mentioned/provided or the user can fill that part with "the reliable guidelines and/or expert consensus on this clinical problem", which is still helpful to generate SCT items.

After these parts are customized based on specific needs, the prompt is ready for using in an LLM environment. We used it to generate SCT items through ChatGPT-4, Claude 3 (Opus), Llama 3 (70B-Instruct), and ChatGPT-4o. In each, we used "undergraduate medical education" as the target group, "diagnosis" as the focus of the questions, and "Right Upper Quadrant Pain" as the clinical problem. We did not mention any specific guideline and left that part as "the reliable guidelines and/or expert consensus on this clinical problem". We used the same prompt in four chatbots.

Table 1. A prompt for generating SCT items in medical education.

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.

Table 2. An SCT item generated by using the prompt in ChatGPT-4.

A 45-year-old male presents with sudden onset of right upper quadrant (RUQ) pain, worsening over the last 6 hours. He has no history of similar symptoms.		
If you were thinking of ...	**Then you learn that ...**	**It becomes ...**
Acute Cholecystitis	The patient's ultrasound shows gallstones with gallbladder wall thickening.	**+2 much more appropriate** (findings support diagnosis)
Acute Hepatitis	Liver function tests show markedly elevated liver enzymes.	**+1 more appropriate** (consistent but not diagnostic)
Peptic Ulcer Disease	The patient reports relief of pain after eating.	**-1 less appropriate** (symptom uncommon for PUD in RUQ)

The SCT item (Table 2) generated by ChatGPT-4 in May 2024 is accessible via this link: https://chat.openai.com/share/ffd61682-eceb-4912-8043-d3d9fb31150c

Since Claude does not have a feature to share the conversations via a link (as of May 2024), we could not share the conversation but we followed the same procedure in the process of generating the SCT item (Table 3) in the conversation (provide the same prompt and get the response, in May 2024).

Table 3. An SCT item generated by using the prompt in Claude 3 (Opus).

A 45-year-old obese woman presents with a 2-day history of severe right upper quadrant pain, nausea, and vomiting after eating a fatty meal.		
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)**
Acute cholecystitis	She has a fever of 38.5°C (101.3°F) and right upper quadrant tenderness on palpation.	+2 much more appropriate. Fever and localized tenderness strongly suggest acute inflammation of the gallbladder.
Peptic ulcer disease	She has no history of NSAID use or Helicobacter pylori infection.	-1 less appropriate. Absence of major risk factors for peptic ulcer makes this diagnosis less likely.
Hepatitis	Her liver enzymes (AST, ALT) are within normal limits.	-2 much less appropriate. Normal liver enzymes argue against hepatitis as the cause of her symptoms.

The SCT item (Table 4) generated by Llama 3 (Meta-Llama-3-70B-Instruct) in May 2024 is accessible via this link: https://hf.co/chat/r/DISaGJw

The SCT item (Table 5) generated by ChatGPT-4o in May 2024 is accessible via this link: https://chat.openai.com/share/8fad1aac-b1f4-4e4a-9a2c-60d08787863d

Table 4. An SCT item generated by using the prompt in Llama 3.

A 35-year-old woman presents to the emergency department with a 2-day history of severe, constant right upper quadrant (RUQ) pain that radiates to her back. She denies fever, nausea, or vomiting.		
If you were thinking of...	**Then you learn that...**	**It becomes...**
Cholelithiasis	The patient has a history of gallstones and has been experiencing similar episodes of RUQ pain in the past.	+2 much more appropriate
Acute Hepatitis	The patient's liver function tests (LFTs) are elevated, with an alanine transaminase (ALT) level of 500 IU/L.	+1 more appropriate
Pyelonephritis	The patient has a urinary tract infection (UTI) diagnosed by urinalysis, but her flank pain is minimal.	-1 less appropriate

Table 5. An SCT item generated by using the prompt in ChatGPT-4o.

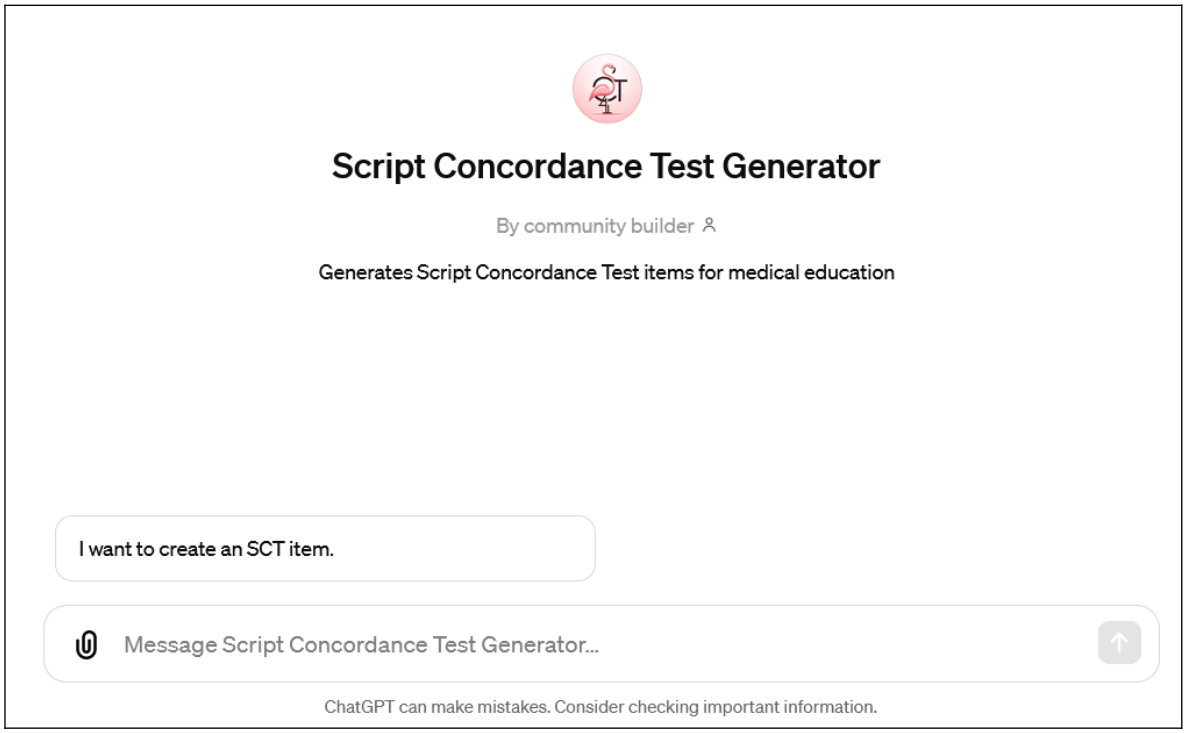
A 45-year-old woman presents to the emergency department with a 12-hour history of severe right upper quadrant (RUQ) pain. She describes the pain as constant and sharp, radiating to her right shoulder. She has had similar episodes in the past, but they were less severe and resolved spontaneously. She denies fever, chills, nausea, or vomiting. Her medical history is significant for hyperlipidemia and obesity. Physical examination reveals tenderness in the RUQ with a positive Murphy's sign.		
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)**
Acute Cholecystitis	The patient has a fever of 101°F (38.3°C) and an elevated white blood cell count.	+2 much more appropriate. Fever and leukocytosis strongly suggest inflammation or infection of the gallbladder.
Biliary Colic	The pain is intermittent and subsides completely after a few hours.	-2 much less appropriate. Biliary colic is characterized by episodic pain that resolves, not constant pain.
Peptic Ulcer Disease	The patient reports taking high-dose NSAIDs for chronic back pain.	+1 more appropriate. NSAID use increases the risk of peptic ulcers, which can cause referred RUQ pain.

In our informal evaluation, we noticed that the generated SCT items appear promising, but with some limitations such as the absence of a detailed Likert scale description. The content seems potentially useful to begin as a foundation. The problems in the LLM-generated SCTs can be corrected through a manual revision by subject matter experts. Such expert review is a requirement, as LLMs are known to, sometimes, generate inaccurate content, and the literature on automatic item generation using LLMs strongly advises against administering such items in the exams without expert review (12,15,19). The current capabilities do not enable us to use them directly. However, it is evident that LLMs can facilitate a streamlined item development process. The ability to generate these SCTs within a matter of seconds is remarkable.

To make the SCT generation process even easier, we also developed a custom GPT (20), named "Script Concordance Test Generator", similar to our Case-based MCQ Generator (21). This free tool eliminates the need for copying and pasting detailed prompts; instead, it only asks users to fill in the necessary parts (figure 1).

The custom GPT is accessible via this link for free (for those who have ChatGPT Plus subscription required by OpenAI, as of April 2024):
https://chat.openai.com/g/g-RlzW5xdc1-script-concordance-test-generator

Figure 1. The custom GPT for Script Concordance Test generation.



Medical education research on this topic can be built upon Hudon et al.'s pioneering work (18). Our prompt and custom GPT provide an opportunity for this purpose. Since our article is limited with providing the prompt and one example SCT item for four LLMs, there still are many research questions that need to be answered, as we mentioned in our previous article on using ChatGPT to generate case-based MCQs (11), such as, assessing the scientific accuracy and clinical relevance of these tests, ensuring the tests meet psychometric standards, and comparing the performance of LLM-generated SCTs against those created by human experts. Additionally, it is important to examine if SCTs can be adapted across various healthcare education settings.

Conclusions

- LLMs are able to generate SCTs in seconds when users benefit from a well-designed prompt.
- Researchers should carry out studies to evaluate the quality of SCTs generated by using LLMs.

Funding: There has been no funding.

Declaration of conflict of interest: The authors declare that there is no conflict of interest.

References

1. ten Cate O. Introduction. In: ten Cate O, Custers EJFM, Durning SJ (eds.) *Principles and Practice of Case-based Clinical Reasoning Education: A Method for Preclinical Students*. Cham: Springer International Publishing; 2018. p. 3–19. https://doi.org/10.1007/978-3-319-64828-6_1.
2. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. *Academic Medicine*. 2019;94(6): 902–912. https://doi.org/10.1097/ACM.0000000000002618.
3. Charlin B, Van Der Vleuten C. Standardized Assessment of Reasoning in Contexts of Uncertainty: The Script Concordance Approach. *Evaluation & the Health Professions*. 2004;27(3): 304–319. https://doi.org/10.1177/0163278704267043.
4. Gheihman G, Johnson M, Simpkin AL. Twelve tips for thriving in the face of clinical uncertainty. *Medical Teacher*. 2020;42(5): 493–499. https://doi.org/10.1080/0142159X.2019.1579308.
5. Moulder G, Harris E, Santhosh L. Teaching the science of uncertainty. *Diagnosis*. 2023;10(1): 13–18. https://doi.org/10.1515/dx-2022-0045.
6. Fournier JP, Demeester A, Charlin B. Script Concordance Tests: Guidelines for Construction. *BMC Medical Informatics and Decision Making*. 2008;8(1): 18. https://doi.org/10.1186/1472-6947-8-18.
7. Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: From theory to practice: AMEE Guide No. 75. *Medical Teacher*. 2013;35(3): 184–193. https://doi.org/10.3109/0142159X.2013.760036.
8. Mathieu S, Couderc M, Glace B, Tournadre A, Malochet-Guinamand S, Pereira B, et al. Construction and utilization of a script concordance test as an assessment tool for dcem3 (5th year) medical students in rheumatology. *BMC Medical Education*. 2013;13(1): 166. https://doi.org/10.1186/1472-6920-13-166.
9. Kün-Darbois JD, Annweiler C, Lerolle N, Lebdai S. Script concordance test acceptability and utility for assessing medical students' clinical reasoning: a user's survey and an institutional prospective evaluation of students' scores. *BMC Medical Education*. 2022;22(1): 277. https://doi.org/10.1186/s12909-022-03339-1.
10. Masters K. Ethical use of artificial intelligence in health professions education: AMEE Guide No.158. *Medical Teacher*. 2023;45(6): 574–584. https://doi.org/10.1080/0142159X.2023.2186203.
11. Kıyak YS. A ChatGPT Prompt for Writing Case-Based Multiple-Choice Questions. *Revista Española de Educación Médica*. 2023;4(3): 98–103. https://doi.org/10.6018/edumed.587451
12. Zuckerman M, Flood R, Tan RJB, Kelp N, Ecker DJ, Menke J, et al. ChatGPT for assessment writing. *Medical Teacher*. 2023;45(11): 1224–1227. https://doi.org/10.1080/0142159X.2023.2249239.
13. Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). Wang J (ed.) *PLOS ONE*. 2023;18(8): e0290691. https://doi.org/10.1371/journal.pone.0290691.
14. Coşkun Ö, Kıyak YS, Budakoğlu İİ. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: A randomized controlled experiment. *Medical Teacher*. 2024; 1–7. https://doi.org/10.1080/0142159X.2024.2327477.
15. Kıyak YS, Coşkun Ö, Budakoğlu İİ, Uluoğlu C. ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *European journal of clinical pharmacology*. 2024;80: 729–735. https://doi.org/10.1007/s00228-024-03649-x.
16. Laupichler MC, Rother JF, Grunwald Kadow IC, Ahmadi S, Raupach T. Large Language Models in Medical Education: Comparing ChatGPT- to Human-Generated Exam Questions. *Academic Medicine*. 2023; https://doi.org/10.1097/ACM.0000000000005626.
17. Kıyak YS, Emekli E. ChatGPT Prompts for Generating Multiple-Choice Questions in Medical Education and Evidence on Their Validity: A Literature Review. *Postgraduate Medical Journal*. 2024. [In-press]
18. 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 Medical Education*. 2024;10: e54067–e54067. https://doi.org/10.2196/54067.

19. Indran IR, Paramanathan P, Gupta N, Mustafa N. Twelve tips to leverage AI for efficient and effective medical question generation: A guide for educators using Chat GPT. *Medical Teacher*. 2023; 1–6. https://doi.org/10.1080/0142159X.2023.2294703.
20. 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. *Medical Teacher*. 2024; 1–5. https://doi.org/10.1080/0142159X.2024.2305365.
21. Kıyak YS, Kononowicz AA. Case-based MCQ generator: A custom ChatGPT based on published prompts in the literature for automatic item generation. *Medical Teacher*. 2024; 1–3. https://doi.org/10.1080/0142159X.2024.2314723.