

IA et création de TCS

JP Fournier

PUPH

Faculté de Médecine – Université Côte d’Azur

Collège National des Enseignants de Thérapeutique



Qu'attendre de l'IA en matière de création de formats d'évaluation ?

- ❖ Gain de temps
- ❖ Gain d'argent
- ❖ Des formats qui « fonctionnent » d'un point de vue métrologique
- ❖ Fournir une rétro-action fiable aux étudiants



D'après

Zuckerman M. *Med Teacher* 2023
Hudon A. *JMIR Med Educ* 2024
Coskun O. *Med Teacher* 2025
Tong L. *Med Teacher* 2025
Lam G. *Med Teacher* 2025

Que tirer de l'expérience de l'IA en matière de QCM ?

- ❖ En thérapeutique :
 - ✓ 15 items créés « traditionnellement » par des spécialistes du domaine (G1) ;
 - ✓ 15 items créés en AIG par le même groupe (G1-AIG) ;
 - ✓ 15 items créés « traditionnellement » par un autre groupe (G2) ;
 - ✓ Evaluation de la qualité des 45 items par 4 experts sur une échelle de likert de 1 à 5 :

Table 2 Mean rating for each of the item quality indicators presented by Haladyna et al.⁶ as a function of test development group

Indicator	Item development group		
	G1, G2 Traditional	G1 AIG	t-test
1 The question measures specific content, as outlined in the test specifications	3.46	3.49	
2 The question is based on important topics in the curriculum and is designed to measure key thinking and problem-solving skills	3.42	3.46	
3 The question is carefully edited, formatted and presented using correct grammar, punctuation, capitalisation and spelling	3.48	3.55	
4 The central idea is included in the stem, not the options	3.62	3.57	
5 The stem of the question is worded positively and avoids negatives such as NOT and EXCEPT	3.69	3.72	
6 Only one of the options is clearly correct	3.27	3.19	
7 The correct option is not cued by item writing errors such as the presentation of a conspicuously correct option or a blatantly incorrect option	3.41	3.31	
8 All of the distractors are plausible	3.54	2.59	p < 0.05*

A higher mean indicates stronger agreement
 * Pairwise comparison
 G1 = group 1; G2 = group 2; AIG = automatic item generation

Que tirer de l'expérience de l'IA en matière de QCM ?

❖ Tout est dans le prompt ! :

You are developing a question bank for medical exams focusing on the topic of [PLEASE INSERT A TOPIC]. Please generate a high-quality single best answer multiple-choice question. Follow the principles of constructing multiple-choice items in medical education. Generate the questions using the following framework:

Case (write as a single narrative paragraph without providing each part separately):

Patient details (gender/age)

Presenting complaint

Relevant clinical history

Physical examination findings

Diagnostic test results (optional)

Question stem: [Insert relevant information from the above sections without compromising the answer]

Acceptable question style: Ask for the BEST answer, NOT one that is TRUE/FALSE.

Answer options:

[Insert plausible answer option]

Explanation:

- Identify and explain the correct answer.
- Explain why this is the most appropriate answer based on evidence-based guidelines or expert consensus.
- Briefly explain why the other answer options are less correct or incorrect.

Difficulty level: [PLEASE INSERT A DIFFICULTY LEVEL (E.G. EASY, DIFFICULT)]

Case:

Mr. Johnson, a 55-year-old male, presents to the primary care clinic with a chief complaint of recurrent headaches. He reports a sedentary lifestyle and a family history of hypertension. His blood pressure readings during previous routine check-ups have consistently been elevated. Mr. Johnson denies any chest pain, shortness of breath, or visual disturbances. His past medical history is significant for obesity and a sedentary occupation. On physical examination, his blood pressure is 150/92 mmHg, and his BMI is 31. Laboratory results show normal renal function and lipid profile.

Question stem:

Given Mr. Johnson's presentation with recurrent headaches, a family history of hypertension, and consistently elevated blood pressure readings, what is the BEST initial management strategy in a primary care setting?

Answer options:

- Initiate lifestyle modifications, including dietary changes and increased physical activity.
- Start antihypertensive medication immediately to lower blood pressure.
- Order a 24-hour ambulatory blood pressure monitoring for further assessment.
- Recommend immediate admission to the hospital for comprehensive evaluation.
- Suggest over-the-counter analgesics for headache relief and reassess in six months.

<https://chat.openai.com/share/4342df6f-7cbe-43df-98dc-b110b53983e3>.

Que tirer de l'expérience de l'IA en matière de QCM ?

- ❖ ChatGPT3.5, ChatGPT4 et Chatpdf
- ❖ Disciplines cliniques et fondamentales (anatomie)
- ❖ Niveau taxonomique peu élevé (1 à 3)
- ❖ QCM de type A, pas systématiquement de vignette clinique
- ❖ Pas d'image (malgré l'utilisation de ChatGPT4)
- ❖ Effectifs très variables (3 à 210 questions)
- ❖ Qualité :
84 à 100% de QCM correctement construits (effectifs très variables)
- ❖ Quand comparés aux QCM conçus par des enseignants (50 QCM) :
 - ✓ Légère supériorité pour les QCM construits par les enseignants, significative pour la pertinence
 - ✓ Aucune différence de qualité pour les étudiants
- ❖ Quand testés :
 - ✓ KR 20 : 0,68
 - ✓ Index de difficulté : $0,48 \pm 0,05$
 - ✓ Index de discrimination : $0,41 \pm 0,08$
 - ✓ Aucune différence entre les QCM générés par IA et ceux préparés par des enseignants

D'après

Zuckerman M. *Med Teacher* 2023
Rezigalla AA. *BMC Med Educ* 2023
Cheung BHH. *PloS One* 2023
Klang E. *BMC Med Educ* 2023
Coskun O. *Med Teacher* 2025

L'IA peut passer des examens comportant des QCM

- ❖ Passer et... réussir des épreuves comportant des QCM
- ❖ Y compris avec des images, théoriquement (ChatGPT 4)
- ❖ Rétro-action possible
- ❖ Examens nationaux multi-disciplinaires
- ❖ Examens de certification de disciplines
- ❖ Environnements culturels variés
- ❖ Mais :
- ✓ Réponses variables en répétant la question : jusqu'à 16 p. cent
- ✓ Pas de réussite dans tous les domaines
- ✓ Pas d'intégration d'images dans les épreuves construites (ChatGPT 4)
- ✓ Principale source d'informations : Internet avec toutes les erreurs qui y sont...
- ✓ Pas de mesure du raisonnement complexe
- ✓ **Nécessité d'un contrôle académique des réponses, et surtout des commentaires**
+++++

D'après

Green ML. *Med Teacher* 2018
Gaube S. *Digital Medicine* 2021
Mbakwe AB. *Plos Digit Health* 2023
Chan J. *Ann R Coll Surg Engl* 2024
Wocik S. *Cardiol J* 2023
Wang H. *Int J Med Inform* 2023
Guerra GA. *World Neurosurg* 2023
Kung TH. *Plos Digital Health* 2023
Tagaki S. *JMIR* 2023
Gilson A. *medRxiv* 2022
Friederichs H. *Med Educ Online* 2023
Meo SA. *Healthcare* 2023
Ayoub Meo S. *Healthcare* 2023
Duong D. *Eur J Hum Genet* 2023
Suchman K. *Am J Gastroenterol* 2023
Wang YM. *J Chin Med Assoc* 2023
Humar P. *Aesthetic Surg* 2023
Huh S. *J Educ Eval Health Prof* 2023
Klang E. *BMC Med Educ* 2023
Temsah MH. *Healthcare* 2023

TCS générés par IA

- ❖ 3 séries publiées :
 - ✓ Psychiatrie : 3 vignettes
 - ✓ Gynéco-obstétrique : 5 vignettes (?)
 - ✓ Imagerie : 5 vignettes
- ❖ Méthodologie similaire :
 - ✓ Confrontation à un panel d'experts qui juge la qualité sur 11 à 12 critères
 - ✓ Dans 1 étude : experts capables de reconnaître l'origine du TCS (IA vs « humain ») et comparaison avec 3 TCS humains pour la qualité
 - ✓ Dans 1 étude : utilisation de 2 LLM
 - ✓ Dans 1 étude : mesure de la variabilité des réponses du panel d'experts
- ❖ Résultats :
 - ✓ Concordance avec les critères de qualité : 68,06 à 90,57 p. cent
 - ✓ Pas de différence de qualité entre TCS générés par IA et par humains (3 TCS)
 - ✓ TCS générés par IA reconnus comme tels par 39,21 p. cent des panelistes (52,94 p. cent pour les TCS humains, $p = \text{NS}$) (3 TCS)
 - ✓ Si plusieurs LLM utilisés : ChatGPT supérieur (1 étude)
 - ✓ TCS générés par IA adaptés, mais plus scolaires, voire caricaturaux (3 TCS)

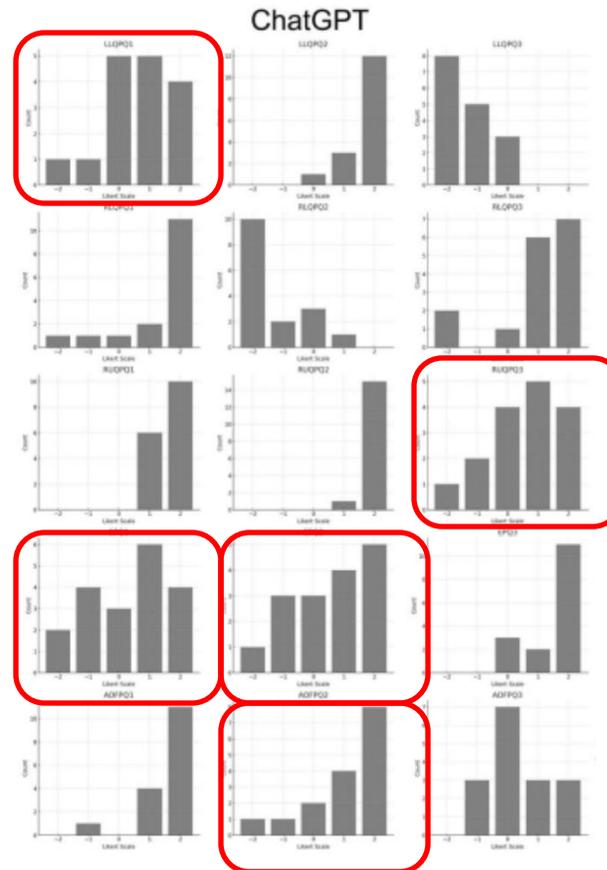
D'après

Hudon A. *JMIR Med Educ* 2024
Coskun ZY. *Med Teacher* 2025
Kriyak YS. *Rev Esp Edu Med* 2025

TCS générés par IA

Niveau de variabilité : 5 vignettes – 15 questions :

❖ 10/15 correctes :



- ❖ Conditionne le pouvoir discriminant de la question
- ❖ Traitement des réponses ou panelistes « déviants » pas établi
- ❖ Pas d'impact sur la reproductibilité des scores si au moins 15 panelistes (TCS de certification)
- ❖ Impact en termes de feedback ?

D'après

Kriyak YS. *Rev Esp Edu Med* 2025
Wan M. *Med Educ* 2018
Cooke S. *Acad Med* 2017
Power A. *Teach Learn Med* 2017
Gagnon R. *Adv Health Sci Educ* 2011

TCS générés par IA

Qui a fait quoi ?

Vous voyez aux urgences une femme de 57 ans qui consulte pour des céphalées rebelles.

Si vous pensiez à...	Et qu'alors vous apprenez que...	L'impact sur votre hypothèse est...				
12.1- Une artérite à cellules géantes	<u>elle</u> rapporte une baisse de la vision de l'œil gauche depuis la veille	-2	-1	0	+1	+2
12.2- Des métastases cérébrales	<u>elle</u> est en cours d'exploration pour des rectorragies	-2	-1	0	+1	+2
12.3- Une thrombophlébite cérébrale	<u>elle</u> prend un traitement hormonal substitutif de la ménopause	-2	-1	0	+1	+2
-1 : Négatif -2 : Fortement négatif	0 : Cela ne change rien	+1 : Positif +2 : Fortement positif				

Réponses attendues :

12.1 : +2
12.2 : +1
12.3 : 0, +1

Commentaires :

C'est le diagnostic à évoquer d'emblée. Il y a urgence à débiter une corticothérapie systémique. La corticothérapie n'empêche pas la réalisation d'une biopsie d'artère temporale. Le risque thrombo-embolique est augmenté en cas de traitement administré per os et non en cas de traitement administré par voie trans-dermique. La nature du progestatif conditionne également en partie le risque thrombo-embolique.

Vous voyez aux urgences une femme de 42 ans qui consulte pour des céphalées diffuses, progressives depuis 10 jours, devenues intenses, continues et rebelles aux antalgiques habituels (paracétamol et ibuprofène). Elle ne rapporte pas de photophobie ni de nausées. Elle a comme antécédent une hypertension artérielle traitée par ramipril. Elle ne prend pas de contraception orale. L'examen neurologique est normal.

Si vous pensiez à une...	Et qu'alors vous apprenez que...	L'impact sur votre hypothèse est...				
5.1- Thrombophlébite cérébrale	<u>L'IRM</u> cérébrale en séquence T1 post-gadolinium montre un comblement du sinus sagittal supérieur	-2	-1	0	+1	+2
5.2- Hypertension <u>intra-crânienne</u> bénigne	<u>le</u> fond d'œil montre un œdème papillaire bilatéral	-2	-1	0	+1	+2
5.3- Hémorragie sous arachnoïdienne	<u>le</u> scanner cérébral sans injection est normal	-2	-1	0	+1	+2
-1 : Négatif -2 : Fortement négatif	0 : Cela ne change rien	+1 : Positif +2 : Fortement positif				

Réponses attendues :

5.1 : +2
5.2 : +2
5.3 : -1

Commentaires :

Un comblement du sinus sagittal supérieur en IRM injectée est un signe caractéristique de thrombose veineuse cérébrale. Cela confirme l'hypothèse diagnostique. La thrombose veineuse cérébrale est une cause classique de céphalées subaiguës, inhabituelles, rebelles et isolées chez une femme jeune.

L'œdème papillaire bilatéral est un élément clé pour évoquer une hypertension intracrânienne. Chez une femme d'âge moyen, en l'absence de signes focaux et avec un fond d'œil anormal, cela renforce nettement cette hypothèse.

Un scanner cérébral sans injection réalisé au-delà de 6 heures et surtout après plusieurs jours de céphalées a une sensibilité diminuée pour détecter une hémorragie méningée. Un scanner normal à ce stade réduit la probabilité mais ne permet pas d'exclure totalement ce diagnostic. Un complément par une ponction lombaire pourrait être discuté si le contexte clinique le justifie.

Construction de TCS par IA

- ❖ Utiliser préférentiellement ChatGPT4
- ❖ Tout est dans le prompt !
- ❖ ...Et la relecture !
- ❖ De façon générale pour les prompts :
- ✓ Clarté
- ✓ Contexte
- ✓ Cohérence avec l'objectif
- ✓ Préciser la forme de la réponse attendue
- ✓ Garde-fous de sécurité (vie privée, etc...)

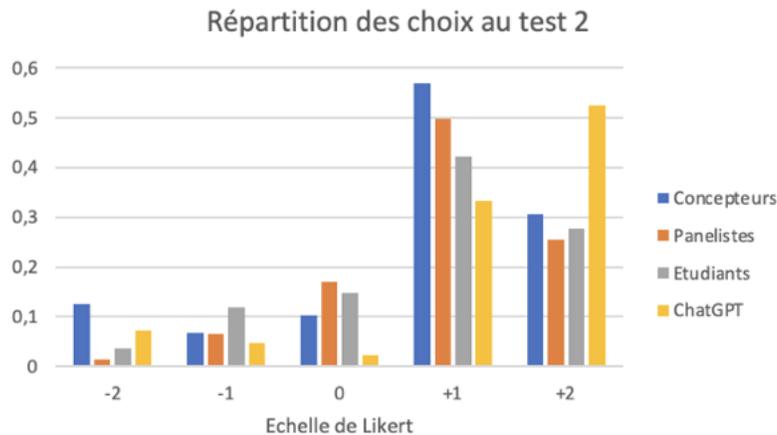
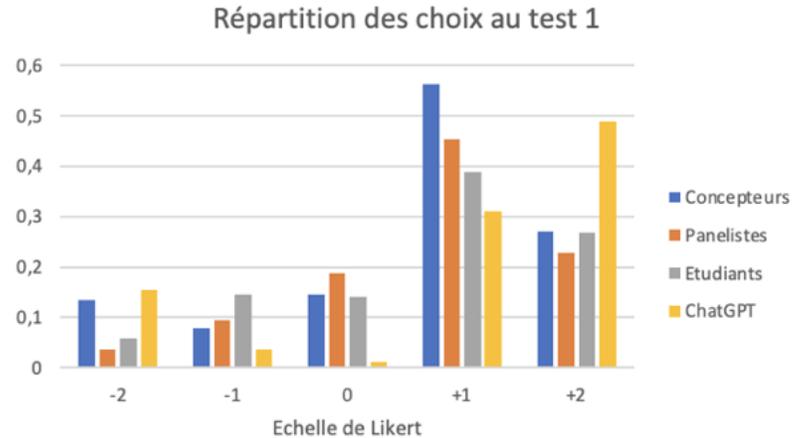
<https://chatgpt.com/g/g-RlzW5xdc1-script-concordance-test-generator>

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

<p>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"].</p> <p>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.</p> <p>Each SCT consists of four main components.</p> <ol style="list-style-type: none"> 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)". <p>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.</p> <p>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.</p> <p>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"</p>



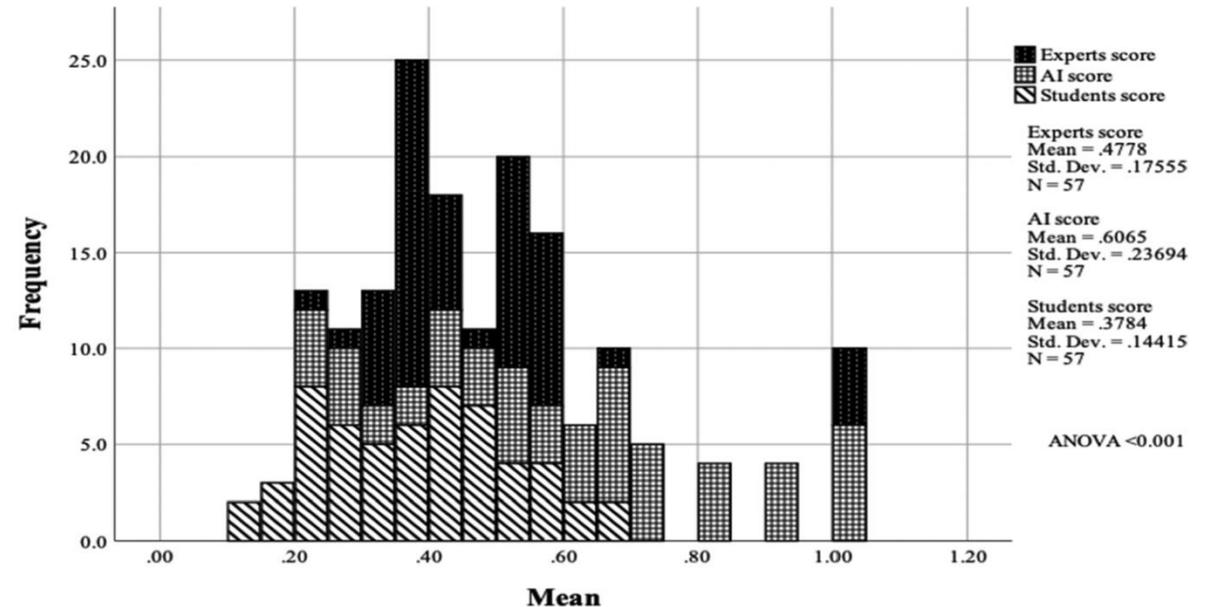
ChatPT peut passer des épreuves de TCS



- ❖ Par rapport aux étudiants, ChatPT favorise les valeurs extrêmes (-2 et surtout +2)
- ❖ L'excès des choix extrêmes (+2 et à un moindre niveau -2) explique les scores légèrement plus faibles :
 - ✓ 55,45/100 (63,74 ± 8,80, p = NS)
 - ✓ 57,85/100 (65,00 ± 9,40, p = NS)
- ❖ 10 vignettes communes aux 2 tests : 51,41 et 54,54/100 (étudiants : 63,77 ± 10,05 et 63,75 ± 11,15, p = NS)

ChatGPT peut générer un panel d'experts

- ❖ En ophtalmologie
- ❖ Rédaction de 17 vignettes et 57 questions
- ❖ Soumission à un panel de 15 experts et construction d'un panel virtuel par ChatGPT4 et o1Preview
- ❖ Comparaison des scores



Etudiants : $0,38 \pm 0,14$

Panelistes ChatGPT : $0,48 \pm 0,18$

Panelistes : $0,61 \pm 0,24$

ChatGPT peut générer un panel d'experts

Oui, mais...

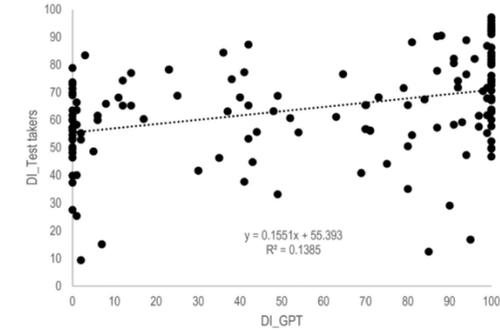
Vous voyez aux urgences une femme de 26 ans qui consulte pour une douleur thoracique droite. Votre principale hypothèse est une embolie pulmonaire sans élément de gravité, d'autant que les D-dimères sont à 1750 ng/mL. Son principal antécédent est une allergie vraie aux produits de contraste iodés (œdème de Quincke au décours d'un scanner abdominal injecté il y a 2 ans). Elle prend aucun traitement régulier en dehors d'une contraception orale.

Si vous pensiez prescrire...	Et qu'alors vous apprenez que...	L'impact sur votre hypothèse est...				
1.1- Un échodoppler veineux des membres inférieurs	<u>elle</u> ne présente aucun signe clinique de thrombose veineuse profonde des membres inférieurs	-2	-1	0	+1	+2
1.2- Une scintigraphie pulmonaire ce ventilation/perfusion	<u>l'examen</u> thoracique ne montre qu'un discret frottement pleural à la base droite	-2	-1	0	+1	+2
1.3- Une tomodensitométrie thoracique non injectée	<u>sa</u> température auriculaire est à 38,2°C	-2	-1	0	+1	+2
-1 : Négatif -2 : Fortement négatif	0 : Cela ne change rien	+1 : Positif +2 : Fortement positif				

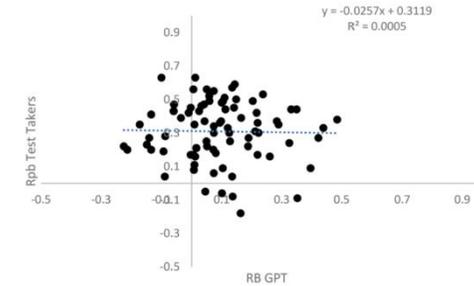
Composition du panel	Question	Réponses attendues	-2	-1	0	+1	+2
15 urgentistes	Q1	-1	6	9	9	4	
	Q2	+2					
	Q3	0, +1					
5 urgentistes PU 5 urgentistes PH 5 urgentistes internes	Q1	-1	8	7	8	4	
	Q2	+2					
	Q3	0, +1					
11 urgentistes 2 pneumologues 2 radiologues	Q1	-1	9	6	10	3	
	Q2	+2					
	Q3	0, +1					
5 urgentistes 5 pneumologues 5 radiologues	Q1	-1	8	7	9	3	
	Q2	+2					
	Q3	0, +1					

ChatGPT peut prédire les qualités psychométriques de questions

- ❖ Examen national classant portugais 2020
- ❖ 150 QCM type *one best answer*
- ❖ Chaque question soumise 100 fois à ChatGPT4.o
- ❖ Corrélation entre les niveaux de difficulté (*p-value*) prédit et observé
- ❖ Corrélation entre les coefficients de corrélation bisérielle (*r-bis*) prédit et observé
- ❖ Corrélation entre les questions défailtantes et le nombre de plaintes



P = 0,001



P = NS

Table 2. Correlation between GPT's prediction and number of challenges.

	Difficulty Index GPT	RB_GPT	Challenges
Difficulty Index GPT	–		
R _{pb} _GPT	0,089	–	
Challenges	–0.302**	–0.020	–

P = 0,001

R_{pb}-Point-Biserial Correlation coefficient.
**Correlation is significant at the 0.01 level.

Messages Clés – *Take-Home Messages*

Twelve tips to leverage AI for efficient and effective medical question generation: A guide for educators using Chat GPT

Inthrani Raja Indran* , Priya Paranthaman*, Neelima Gupta  and Nurulhuda Mustafa

Department of Pharmacology, National University of Singapore, Yong Loo Lin School of Medicine, Singapore, Singapore

ABSTRACT

Background: Crafting quality assessment questions in medical education is a crucial yet time-consuming, expertise-driven undertaking that calls for innovative solutions. Large language models (LLMs), such as ChatGPT (Chat Generative Pre-Trained Transformer), present a promising yet underexplored avenue for such innovations.

Aims: This study explores the utility of ChatGPT to generate diverse, high-quality medical questions, focusing on multiple-choice questions (MCQs) as an illustrative example, to increase educator's productivity and enable self-directed learning for students.

Description: Leveraging 12 strategies, we demonstrate how ChatGPT can be effectively used to generate assessment questions aligned with Bloom's taxonomy and core knowledge domains while promoting best practices in assessment design.

Conclusion: Integrating LLM tools like ChatGPT into generating medical assessment questions like MCQs augments but does not replace human expertise. With continual instruction refinement, AI can produce high-standard questions. Yet, the onus of ensuring ultimate quality and accuracy remains with subject matter experts, affirming the irreplaceable value of human involvement in the artificial intelligence-driven education paradigm.

ARTICLE HISTORY

Received 4 July 2023

Accepted 11 December 2023

KEYWORDS

Chat GPT; AI; medical assessment; questions