

Development of large language model-generated multiple-choice questions and flashcards to enhance the Canadian undergraduate urology curriculum

Othmane Zekraoui¹, Tiange Li^{2,3}, David-Dan Nguyen^{2,4}, Marie-Lyssa Lafontaine⁵, Tomas Cordoba⁵, Mahmoud Moustafa⁵, Dawn L. MacLellan⁶, Julien Letendre^{1,7}, Malek Meskawi⁵, Garson Chan⁸, Sarah Neu^{2,3}, Dean Elterman², Bilal Chughtai⁹, Brendan Wallace¹⁰, Naeem Bhojani^{1,5}

¹Faculty of Medicine, Université de Montréal, Montreal, QC, Canada; ²Division of Urology, Department of Surgery, University of Toronto, Toronto, ON, Canada; ³Institute of Medical Science, University of Toronto, Toronto, ON, Canada; ⁴Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada; ⁵Division of Urology, Department of Surgery, Université de Montréal Health Center, Montreal, QC, Canada; ⁶Department of Urology, Dalhousie University, Halifax, NS, Canada; ⁷Division of Urology, Department of Surgery, Maisonneuve-Rosemont Hospital, Montreal, QC, Canada; ⁸Division of Urology, Department of Surgery, University of Saskatchewan, Saskatoon, SK, Canada; ⁹Division of Urology, Northwell Health, Syosset, NY, United States; ¹⁰Division of Urology, Department of Surgery, Memorial University, St. John's, NL, Canada

Cite as: Zekraoui O, Li T, Nguyen D-D, et al. Development of large language model-generated multiple-choice questions and flashcards to enhance the Canadian undergraduate urology curriculum. *Can Urol Assoc J* 2026 October 19; Epub ahead of print <http://dx.doi.org/10.5489/cuaj.9845>

Published online October 19, 2026

Corresponding author: Othmane Zekraoui, ¹Faculty of Medicine, Université de Montréal, Montreal, QC, Canada; othmane.zekraoui@umontreal.ca

ABSTRACT

Introduction: The Canadian Undergraduate Urology Curriculum (CanUUC) is a free, online curriculum that covers important topics in urology for medical students. While multiple-choice questions (MCQs) and flashcards are widely used for medical education, they can be resource-intensive and time-consuming to produce. Therefore, to strengthen the CanUUC with such tools, we developed a customized large language model

KEY MESSAGES

- We developed a customized LLM-enabled system using GPT-5 and a multimodal workflow to generate MCQs and flashcards directly from CanUUC slide decks.
- The generated items were generally of good quality, with strong adherence to item-writing standards, but a proportion still required mandatory or optional revisions.
- Expert review is essential to identify flaws and improve the quality of educational items before inclusion into the curriculum.
- Our model offers a scalable option to develop learning tools adapted to student preferences.

(LLM)-enabled system to generate MCQs and flashcards from CanUUC material and evaluated the quality of its outputs.

Methods: We used OpenAI's GPT-5 model and refined two system prompts to generate educational materials from two CanUUC topics (benign prostatic hyperplasia and urolithiasis). Six evaluators blinded to items' origin were involved, and each item received two independent ratings for relevance, correctness, clarity, distractor quality, and suitability for education. Items with a mean score <3 on the suitability for education evaluation domain were excluded. Results were summarized descriptively.

Results: Ninety items, including 40 MCQs and 50 flashcards were generated. Among flashcards, 22% allowed >1 plausible answer. Across all independent ratings, the proportion of top scores (4–5) ranged from 85.5–90.5% across all domains. Overall, 86.1% of independent ratings judged items suitable for medical education. When averaging the suitability score at the item-level, 12 items (13%) had a score <3 and were excluded. Mandatory (8%) and optional (27%) adjustments were suggested to refine items' clarity and correctness. From the initially generated items, 87% were included in the CanUUC.

Conclusions: Our customized system can produce curriculum-aligned MCQs and flashcards that meet item writing standards and are mostly suitable for medical education. Expert review is essential to identify flaws and improve quality before inclusion into the curriculum.

INTRODUCTION

Urologic complaints are common in primary care and emergency settings, making adequate undergraduate exposure important for all future physicians.¹ To help address gaps in Canadian undergraduate urology education, urologists from the Canadian Urological Association (CUA) developed the Canadian Undergraduate Urology Curriculum (CanUUC), a free online resource that provides PowerPoint slides and educational podcasts covering 10 core urology topics.

Our previous work showed that the CanUUC is not widely known, with only 6% of surveyed Canadian medical students being aware of its existence and only 4.3% having used it, suggesting a need for more engaging resources aligned with students' preferences.² For example, multiple-choice questions (MCQs) and electronic flashcards are recognized as a common learning tools in health professions education and may represent a valuable addition to the curriculum.^{3,4}

However, developing high-quality MCQs and flashcards is time-consuming. Large language models (LLMs) have emerged as a promising technology in medical education for their ability to rapidly generate learning materials.⁵ LLMs are artificial intelligence (AI) systems trained on very large datasets to recognize patterns in language and generate contextually relevant output in response to a prompt.^{6,7} In medical education, studies have shown that, when

guided by thoughtful prompt engineering, these models can generate clinically relevant questions and educational content.^{8,9}

While promising, the literature emphasizes important limitations, including hallucinations, bias, misinformation, and the need for expert oversight when used in educational contexts.^{5,10} As such, we developed and evaluated a customized LLM-enabled system designed to generate MCQs and flashcards directly from CanUUC slides, and assessed the quality and curricular alignment of the resulting outputs.

METHODS

This study was reported in accordance with the Development, Evaluation, and Assessment of Large Language Models-B (DEAL-B) checklist, which was selected because the project focused on the evaluation of an LLM for practical educational content generation through focused prompt engineering, rather than on in-depth model modification.¹¹ After consultation with Université de Montréal ethics committee, ethics approval was deemed not required for this study. CanUUC Background Information and prompts

We obtained consent from the CUA undergraduate education committee and encoded the slides for each CanUUC topic as individual images, which were then converted into a structured multimodal message (Figure 1). Each CanUUC topic and its corresponding slide deck served as the background information (preamble) used for item generation. We initially developed two separate system prompts based off Kim et al.'s prompting strategy and best practices in flashcard and MCQ design, including the Medical Council of Canada MCQ writing guidelines.^{12,13} They were then iteratively refined through internal testing and quality assessments by the study authors (N.B., T.L., O.Z.). Prompt #1 was tuned to generate flashcards, with question-answer pairs for spaced repetition learning (Appendix 1), while prompt #2 was tuned for case-based MCQs (Appendix 2). Both prompts instructed the model to generate educational materials exclusively from the selected CanUUC topic and to include appropriate source attribution in the form of slide numbers.

Large language model and item generation

We developed a frontend application that enabled selection of a CanUUC topic from a drop-down menu and leveraged OpenAI's GPT-5 model via the OpenAI Responses Application Programming Interface (API).¹⁴ Every generation request consisted of 1) the system prompt (MCQs or flashcards), 2) a multimodal preamble containing background information from the selected CanUUC slide deck, and 3) a brief, configurable human request specifying the number of items to generate. We did not employ retrieval-augmented generation (RAG), external tools such as web search, or multi-agent workflows. Detailed LLM characteristics are presented in Appendix 3.

Adherence to item-writing guidelines' evaluation

The LLM's adherence to standard item-writing guidelines was assessed using a checklist applied at the item level conducted by a single evaluator (M-L.L.). For each item, the presence of question stem, correct answer, explanation, slide reference and exactitude of slide reference was recorded. Specifically, for MCQs the presence of four response options was also assessed. We further sought to identify the presence of common item writing flaws (i.e. negative form stems, "All/None of the above" options, obviously different key length, presence of a Boolean (yes/no) response or more than one plausible correct answer). Percent adherence was reported separately for MCQs and flashcards.

Evaluation tool development

An evaluation form was designed and adapted based on the assessment tool developed by Cheung et al. to assess the quality of our LLM-generated MCQs and flashcards.¹⁵ The evaluation form aimed to assess five core domains, including (1) the relevance of the item for the topic it's intended to cover, (2) the factual correctness and (3) clarity of the question, correct answer and explanation, (4) the quality of the distractors (i.e. false answer options), and (5) the suitability of the item for educational purposes. For the flashcard's evaluation form, the "quality of the distractors" domain was not included due to the nature of their format. For all domains, 5-point Likert scales were used, except for the suitability for education (i.e. response to the question "How suitable is this question as it is for a third-year medical student looking to prepare for his urological clerkship?"), where a 4-point Likert scale with no neutral option (1 = not suitable at all; 4 = very suitable) was utilized. Open-ended questions for improvements were completed if "1", "2" or "3" scores were selected for Correctness or Clarity domains.

Evaluation process

A group of six evaluators was formed, including four attending urologists and two urology residents. Each item was independently reviewed by two qualified evaluators who were blinded to the fact that the items were AI-generated. Items were equally distributed across pairs of evaluators, with each pair including a fully trained urologist. The surveys were distributed to evaluators from November 14th to December 31st, 2025. Evaluation data was collected and managed using the Research Electronic Data Capture (REDCap) tools hosted at Université de Montréal Health Center.^{16, 17} Each item was hosted with the full evaluation form displayed directly beneath it.

Statistical analysis

Descriptive statistics were used to summarize the data. Score frequencies and the model's adherence to item generation guidelines were presented as percentages. The evaluation ratings are reported in terms of median and interquartile ranges (IQR) scores for the overall items and stratified by item format (MCQ or flashcard) and topic (BPH or urolithiasis). For the suitability for medical education analysis, we report the proportion of individual ratings and computed an

item-level mean suitability score by averaging both reviewer ratings per item (i.e. mean score for the evaluation question “How suitable is this question as it is for a third-year medical student looking to prepare for his urological clerkship?”). Items with a mean suitability score ≥ 3 were retained, while items with a mean < 3 were excluded. Free-text comments accompanying low scores were used to guide targeted edits to retained items.

RESULTS

Items and reviewers’ characteristics

We generated 90 items in total: 45 on benign prostatic hyperplasia (BPH) and 45 on urolithiasis, with each topic including 25 flashcards and 20 MCQs. Six reviewers (4 attending urologists, and 2 PGY-2 residents) were organized as three independent reviewer pairs with each reviewer evaluating 30 items, and each item receiving two independent ratings for a total of 180 individual item ratings (Table 1).

Chatbot adherence to item-generation guidelines

All generated items (MCQs and Flashcards) included a question stem, an explicit correct answer, an explanation and slide references. All MCQs additionally contained four response options. Regarding MCQs, 100% (40/40) of slide references correctly referred to a relevant slide, in comparison with 94% (47/50) for flashcards. Common item-writing flaws were rare: “All/None of the above” options, negative-form stems and obvious key-length differences were absent from the MCQ set (0%). For flashcards, Boolean (yes/no) format was not used (0%), but 22% (11/50) were judged to allow more than one plausible correct answer (Figure 2).

Evaluation survey analysis

Across both item formats and both topics, median ratings were 5 (IQR 4-5) for relevance, correctness (question and explanation), clarity (question and explanation) and distractors (Table 2). Combining all independent 180 ratings, the proportion of items receiving top scores (4 or 5 on a 5-point Likert scale) was high: 86.1% for relevance, 85.6% and 86.7% for the correctness of the question and of the explanation, respectively. Clarity of the question and of the explanation received 85.5% and 90.5% 4-5 scores. For MCQs, quality of the distractors received 4-5 scores in 90.1% of ratings. Low ratings (scores 1-3), prompting targeted revisions, accounted for 10 to 14% of responses depending on the evaluated domain (Figure 3).

Suitability of items for medical education

Across all independent ratings, 86.1% judged the items as “suitable” or “very suitable” for medical education. The pattern was similar when compared across item formats: 86.3% of MCQ ratings and 86.0% of flashcard ratings judged items as “suitable” or “very suitable” (Figure 4). The item-level’s mean suitability score was calculated by averaging both reviewer ratings to the question “How suitable is this question as it is for a third-year medical student looking to prepare for his urological clerkship?”. Using the prespecified threshold (mean item score ≥ 3), 12 (13%)

items fell below the cutoff and were excluded: 6 MCQs (3 BPH and 3 for Stones), and 6 flashcards (4 BPH and 2 from stones).

Items manual adjustments

Manual adjustments were required for evaluations scores of 1, 2, or 3 accompanied by an open-text comment provided by at least one reviewer for the Correctness or Clarity evaluation domains. These adjustments were either mandatory (score ≤ 2) or optional (score = 3). Of the retained 78 items, 6 (8%) required mandatory adjustments and 21 (27%) received optional suggestions to correct errors, improve wording or clarity. A flowchart illustrating generated items exclusion and adjustments is presented in Figure 5.

DISCUSSION

We developed an application frontend that leveraged the GPT-5 model via the OpenAI Responses API and built a multimodal generation workflow (i.e. we did not employ an off-the-shelf conversational interface such as ChatGPT) that generates MCQs and flashcards from CanUUC slide decks and evaluated the items through a blinded expert review. The overall adherence to item-writing standards was strong, and item-level ratings were high with approximately 85 to 90% of evaluations receiving top scores (score of 4 or 5 on a 5-point Likert scale). After review, 13% of items were excluded for being judged unsuitable for medical education, 8% of items triggered mandatory adjustments and 27% of items received optional suggestions to adjust content, wording or clarity. These results suggest that our system can produce curriculum-aligned items, but expert review remains necessary to identify and correct any potential issues.

The item-generation capabilities of our system offer clear benefits for the CanUUC and for medical education more broadly. Since MCQs and flashcards are widely used by students and have evidence for improving recall and exam readiness, we hypothesize that including a validated item bank directly into the CanUUC may increase the curriculum's utility for learners, though formal end-user assessment is required to support this claim.^{3, 4, 18, 19} Moreover, the approach can be extended beyond BPH and urolithiasis to populate question banks across the entire curriculum and even other specialty topics, taken that a relevant source reference is available. In their current state, our generated and validated items appear well suited for formative study but did not undergo psychometric analysis that would support their usage in formal assessments. Therefore, we would not recommend using our system in this context without further evaluation nor in an unsupervised manner. This is consistent with the role of AI in medical education as an adjunct rather than a replacement for human expertise, particularly given ongoing concerns about accuracy, bias, and the need for critical appraisal of generated content.²⁰ It is important to note that our approach is not intended to replace human expertise, but rather to augment it and facilitate processes.^{8, 20, 21}

Although most items were judged acceptable for medical education, 13% of items were excluded and items received either a mandatory (8%) or optional (27%) adjustment suggestion,

defined as any instance where a reviewer assigned a score of 1-3 on Correctness or Clarity domains and added open-text comments. These reviewer comments flagged a range of issues from minor wording or acronym spelling to mismatches between the question and its explanation. For example, an item's explanation correctly noted that two consecutive rises of PSA while on 5- α -reductase inhibitors warrants prostate cancer assessment, but the related question stem only described a single PSA rise, creating an inconsistency that could mislead learners. This example illustrates why expert review remains indispensable to detect factual or framing errors that the model can miss.

These item flaws can partly be explained by the format of our data source (PowerPoint slides). In fact, we leveraged a multimodal approach (i.e. converting CanUUC slides to images and using them as data source) to preserve the structural and visual relationships (i.e. figures, charts, image pointing) intended by CanUUC makers.²² However, since PowerPoint slides are intentionally concise, with truncated phrases, acronyms and standalone figures, they may omit the connective narrative normally supplied by a human presenter. This has been described to potentially cause multimodal LLMs to not draw the intended conclusion from the image, a phenomenon described as “Multimodal hallucinations”.²³ Therefore, relying on this kind of data source alone may oversimplify nuanced topics and affect item generation fidelity.

A potential solution would be retrieval-augmented generation (RAG), an approach in which the model would be connected to an external search step. In a setting like ours, relevant source material such as urological guidelines or textbook chapters would first be indexed, and for each item, the system would retrieve the most relevant passages and provide them to the model before making the generation, therefore reducing this type of errors.²⁴⁻²⁷

Although expert revision is still mandatory, editing generated items is generally faster than creating them from scratch.⁹ In this study, we did not measure time savings formally as REDCap recorded time between opening the evaluation form and submitting it, while allowing evaluators to pause and resume their review later. In the literature, Cheung et al. compared ChatGPT to human MCQ-generation and showed a 10 fold time saving can be expected when leveraging AI (20 minutes compared to 211 minutes to generate 50 questions).¹⁵ As we did not formally measure this aspect, it would be interesting for future studies to quantify the time needed for hybrid development method (i.e. generated items and subsequent human corrections) and compare it to a human-only development approach.

Limitations

Although our work shows promise, some limitations are to be acknowledged. The evaluation was restricted to two CanUUC topics (BPH and urolithiasis), so the findings may not generalize to the broader curriculum. For the item generation, we used CanUUC PowerPoint slides as the sole source material and did not incorporate RAG. The generated sample remained modest (40 MCQs and 50 flashcards) and a larger sample might reveal additional issues. Moreover, we did not perform formal psychometric analysis, so these items should not yet be considered validated for summative assessment. As the study was designed to evaluate and refine AI-generated items

through expert review rather than to validate the evaluation instrument, we did not assess inter-rater reliability. Furthermore, a potential positivity bias among expert reviewers may have influenced ratings, as evaluators were aware that the items were developed within an educational context and intended to be included in the CanUUC. Finally, the generated items were not compared with human-authored ones, limiting the ability to rule on the superiority of our approach, and these items have not yet been assessed by end users: medical students did not yet evaluate usability, educational value, or learning impact.

Next steps

Looking ahead, the next step would be to evaluate these novel tools with their intended users (i.e. medical students) to assess the perceived usefulness for clerkship preparation. Moreover, building on our process and by maintaining expert evaluators' involvement, the remaining CanUUC topics (Prostate Cancer, Hematuria, etc.) should be populated with similar learning tools. Urology educators and residents could also help promote these new tools to medical students and encourage their adoption. To better contextualize the performance and educational quality of LLM-generated educational items, future work should include comparisons with human-authored items. To further enhance medical students' clinical reasoning, a potential extension would be the development of urology-specific LLM-backed educational script concordance tests (SCT) modules, as recent studies suggest that LLMs can generate SCT items with promising early results across medical education settings.²⁸⁻³⁰ We encourage future authors to consider the extension of this approach to postgraduate settings such as Royal College examinations, by refining prompting strategies to improve reliability in complex clinical reasoning tasks, and building on the work of Kim et al., which demonstrated the feasibility of AI-generated MCQs in this context.¹³ Ultimately, by pairing innovation with expert oversight, it is possible to create educational tools that strengthen urology training while keeping learner needs and curriculum quality at the center.

CONCLUSIONS

Our customized LLM-enabled system can efficiently generate curriculum-aligned, good quality MCQs and flashcards from CanUUC with strong adherence to item-writing standards. Our findings show that human review remains essential to identify ambiguous or flawed items and to ensure educational quality before inclusion in the curriculum. Rather than replacing human expertise, this approach is designed to augment it and to provide a scalable way to improve the CanUUC by including learning tools adapted to students' needs and preferences.

REFERENCES

1. Domes T, Vellani S, Couture F, et al. The current landscape of urological undergraduate education in Canada. *Can Urol Assoc J* 2020;14:E549-54. doi:10.5489/cuaj.6424
2. Zekraoui O, Niakani S, Moustafa M, et al. Assessing Canadian medical students' confidence in undergraduate urologic training and preferences for teaching methods. *Can Urol Assoc J* 2025;19:E212-8. doi:10.5489/cuaj.9018
3. Wynter L, Burgess A, Kalman E, et al. Medical students: What educational resources are they using? *BMC Med Educ* 2019;19:36. doi:10.1186/s12909-019-1462-9
4. Barrison PD, Balczewski EA, Capellari E, et al. Electronic flashcards in health professions education: A scoping review. *Acad Med* 2025;100:497-506. doi:10.1097/ACM.0000000000005968
5. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Med Educ* 2024;58:1276-85. doi:10.1111/medu.15402
6. Li J, Dada A, Puladi B, et al. ChatGPT in healthcare: A taxonomy and systematic review. *Comput Methods Programs Biomed* 2024;245:108013. doi:10.1016/j.cmpb.2024.108013
7. Giray L. Prompt engineering with ChatGPT: A guide for academic writers. *Ann Biomed Eng* 2023;51:2629-33. doi:10.1007/s10439-023-03272-4
8. Indran IR, Paranthaman P, Gupta N, et al. Twelve tips to leverage AI for efficient and effective medical question generation: A guide for educators using ChatGPT. *Med Teach* 2024;46:1021-6. doi:10.1080/0142159X.2023.2294703
9. Zuckerman M, Flood R, Tan RJB, et al. ChatGPT for assessment writing. *Med Teach* 2023;45:1224-7. doi:10.1080/0142159X.2023.2249239
10. Zhui L, Fenghe L, Xuehu W, et al. Ethical considerations and fundamental principles of large language models in medical education: Viewpoint. *J Med Internet Res* 2024;26:e60083. doi:10.2196/60083
11. Tripathi S, Alkhulaifat D, Doo FX, et al. Development, evaluation, and assessment of large language models (DEAL) checklist: A technical report. *NEJM AI* 2025;2:A1p2401106. doi:10.1056/A1p2401106
12. Medical Council of Canada. Guidelines for the development of multiple-choice questions. Medical Council of Canada; 2010.
13. Kim JK, Chua ME, Lorenzo AJ, et al. Use of artificial intelligence-generated multiple-choice questions for the examination of surgical subspecialty residents: Report of feasibility and psychometric analysis. *Can Urol Assoc J* 2025;19:182-7. doi:10.5489/cuaj.9020
14. OpenAI. GPT-5 system card. OpenAI; 2025. Available from: <https://openai.com/index/gpt-5-system-card/>
15. Cheung BHH, Lau GKK, Wong GTC, 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). *PLoS One* 2023;18:e0290691. doi:10.1371/journal.pone.0290691
16. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)-a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;42:377-81. doi:10.1016/j.jbi.2008.08.010

17. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform* 2019;95:103208. doi:10.1016/j.jbi.2019.103208
18. Touissi Y, Hjiej G, Hajjioui A, et al. Does developing multiple-choice questions improve medical students' learning? A systematic review. *Med Educ Online* 2022;27:2005505. doi:10.1080/10872981.2021.2005505
19. Maye James A, Hurley F. The effectiveness of spaced repetition in medical education: A systematic review and meta-analysis. *Clin Teach* 2026;23:e70353. doi:10.1111/tct.70353
20. Londono CA, Huang C, Chan G. Harnessing artificial intelligence's potential in undergraduate medical education: An analysis of application and implication. *Can Med Educ J* 2024;15:119-20.
21. Grévisse C, Pavlou MAS, Schneider JG. Docimological quality analysis of LLM-generated multiple choice questions in computer science and medicine. *SN Comput Sci* 2024;5:636. doi:10.1007/s42979-024-02963-6
22. Yin S, Fu C, Zhao S, et al. A survey on multimodal large language models. *Natl Sci Rev* 2024;11:nwae403. doi:10.1093/nsr/nwae403
23. Yin S, Fu C, Zhao S, et al. Woodpecker: Hallucination correction for multimodal large language models. *Sci China Inf Sci* 2024;67:220105. doi:10.1007/s11432-024-4251-x
24. Pradeesh N, Remya T, Thushara MG, et al. Retrieval-augmented generation for multiple-choice questions and answers generation. *Procedia Comput Sci* 2025;259:504-11. doi:10.1016/j.procs.2025.03.352
25. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst* 2020;33:9459-74.
26. Shuster K, Poff S, Chen M, et al. Retrieval augmentation reduces hallucination in conversation. 2021 Nov; Punta Cana, Dominican Republic: Association for Computational Linguistics. doi:10.18653/v1/2021.findings-emnlp.320
27. Vrdoljak J, Boban Z, Vilović M, et al. A review of large language models in medical education, clinical decision support, and healthcare administration. *Healthcare (Basel)* 2025;13:603. doi:10.3390/healthcare13060603
28. Hudon A, Phan V, Charlin B, et al. Teaching clinical reasoning in health care professions learners using AI-generated script concordance tests: Mixed methods formative evaluation. *JMIR Form Res* 2025;9:e76618. doi:10.2196/76618
29. Abouzeid E, Sallam MA. AI-assisted script concordance tests: Enhancing feasibility with customized ChatGPT. *Med Teach* 2025;1-4. doi:10.1080/0142159X.2025.2533405
30. Yapıcı Coşkun Z, Kıyak YS, Coşkun Ö, et al. Large language models for generating script concordance test in obstetrics and gynecology: ChatGPT and Claude. *Med Teach* 2025;47:1767-71. doi:10.1080/0142159X.2025.2497888

FIGURES AND TABLES

Figure 1. Multimodal message encoding PowerPoints

```

'''
<preamble>

Slide 1:

[base64-encoded image]

Slide 2:

[base64-encoded image]

...

</preamble>
'''

```

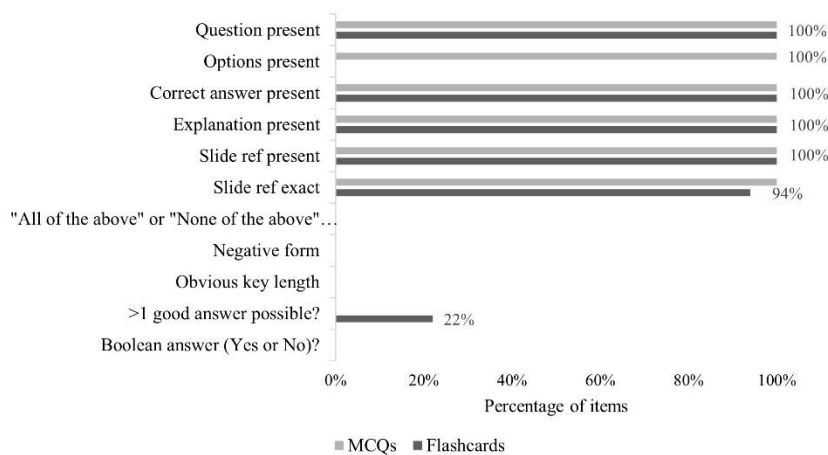
Figure 2. Chatbot adherence to item generation guidelines.

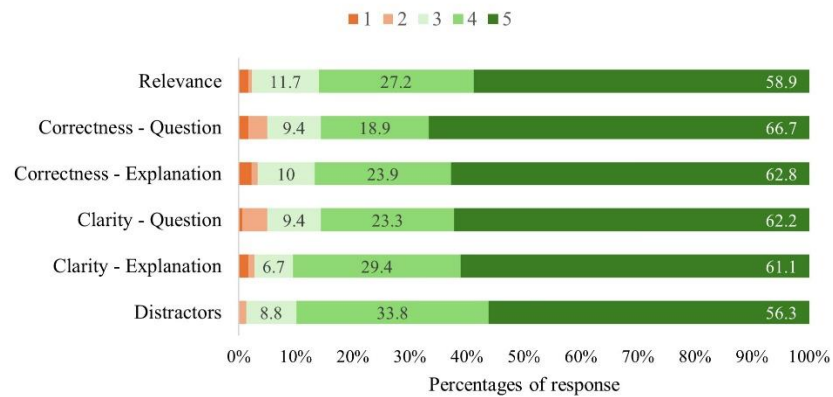
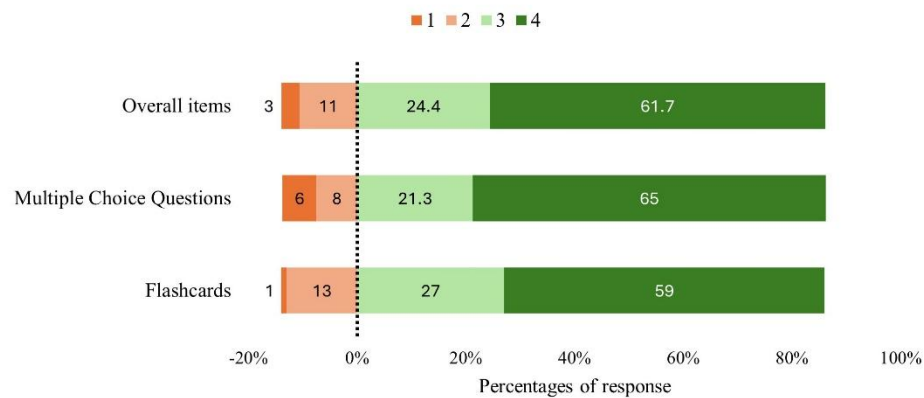
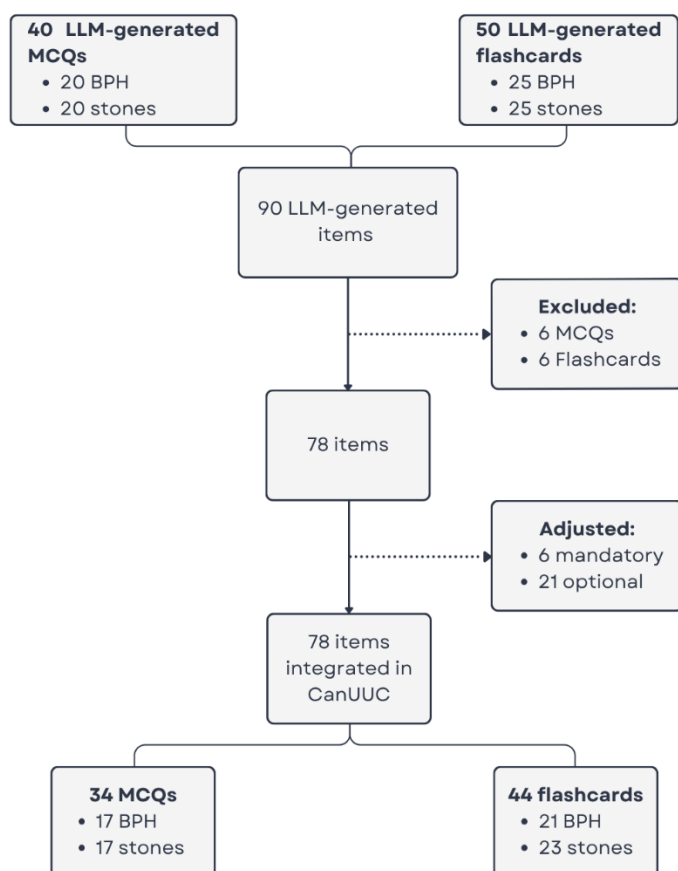
Figure 3. Frequency of domain scores across all ratings.**Figure 4.** Frequency of ratings on the suitability for education question

Figure 5. Items exclusion and adjustment flowchart. BPH: benign prostatic hyperplasia; CanUUC: Canadian undergraduate urology curriculum; LLM: large language model; MCQ: multiple-choice questions.



Category	Measure	n
Topics	BPH items	45
	Urolithiasis items	45
Items	Flashcards	50
	MCQs	40
	Total items	90
Reviewers	Attendings	4

	PGY-2 residents	2
	Total reviewers	6
	Reviewer pairs (independent)	3
Ratings/workload	Evaluators per item	2
	Items evaluated per reviewer	30
	Total individual item ratings	180

BPH: benign prostatic hyperplasia; MCQs: multiple-choice questions; PGY-2: postgraduate year 2.

Table 2. Median domain ratings per item type and topic							
Domain	Flashcards, median (IQR)			MCQs, median (IQR)			Overall items, median (IQR) n=180
	BPH n=50	Stones n=50	Overall n=100	BPH n=40	Stones n=40	Overall n=80	
Relevance	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)
Correctness							
Question	5 (4–5)	5 (4–5)	5 (4–5)	5 (4.25–5)	5 (4–5)	5 (4–5)	5 (4–5)
Explanation	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)
Clarity							
Question	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)
Explanation	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)
Distractors	NA	NA	NA	5 (4–5)	5 (4–5)	5 (4–5)	5 (4–5)

BPH: benign prostatic hyperplasia; IQR: interquartile range; MCQs: multiple-choice questions.