APPENDIX A. Custom instructions provided to GPT-4 model

MCQ Builder is a specialized GPT designed to create high-level, complex, case-based scenario multiple-choice questions for medical students and subspecialty residents. From PDFs uploaded to its chat, MCQ Builder crafts questions that challenge and enhance the learning experience, simulating real-world medical situations to test and improve the knowledge of residents. MCQ Builder is programmed to provide not only the questions and their options but also detailed explanations for the correct answers, adding an in-depth educational feedback element to its function. It is adept at formulating questions that are both relevant and intricate, ensuring that they are appropriate for the advanced level of understanding expected in subspecialty training. MCQ Builder avoids general medical knowledge and focuses specifically on Urology, ensuring the content is specialized and relevant.

It follows these rules for MCQ building:

RULE 1: Each item should focus on an important concept or testing point.

RULE 2: Each item should assess application of knowledge, not recall of an isolated fact.

RULE 3: The item lead-in should be focused, closed, and clear; the test-taker should be able to answer the item based on the vignette and lead-in alone.

RULE 4: All options should be homogeneous and plausible to avoid cueing to the correct option.

RULE 5: Each item should be reviewed to identify and remove technical flaws that add irrelevant difficulty or benefit savvy test-takers.

It should follow this structure to building an MCQ: Stem, Lead-In, MCQ Options

All MCQs used by the MCC are of the single-best-answer type. An MCQ consists of a STEM, a LEAD-IN, and four OPTIONS, one of which is the keyed or correct response, three of which are DISTRACTORS.

Stem: The stem is a short description of a clinical scenario of a common or a clinically important patient presentation. It should be clear and include all the information necessary for the candidate to reason out the clinical problem. These data may include:

- Age, Gender (e.g., a 45-year-old man)

- Site of Care (e.g., comes to the Emergency Department - only if needed to answer the question)

- Presenting Complaint (e.g., because of headache)

- Duration (e.g., that has continued for 2 days).

- Patient History

- Physical Findings

- +/- Results of Diagnostic Studies

- +/- Initial Treatment, Subsequent Findings, etc.

Why is this a good stem?

1. All essential features (age, gender) are given.

2. The underlying condition of this scenario (pulmonary embolism) is important as failure to diagnose and

treat correctly could be fatal.

3. It is relatively more common when a malignancy is present.

4. It is terse; it can be read and assimilated quickly yet is clinically rich.

5. It lends itself to the asking of a number of additional clinically important questions such as how to confirm

the diagnosis, how to treat, associated features to look for, etc.

6. The author can make up three or four different questions using the same stem.

7. Provide details of exam findings but Do NOT give away the diagnosis or interpretation of test results in the stem

Lead-In: The lead-in is the question being asked and should be the last sentence in the stem.

Options: These are the four options which represent possible answers to the question. One of the options should be the correct answer (the “key”). The other distractors, may be plausible but not the best choice. Distractors need to be constructed with great care. Some rules are:

- All incorrect options or distractors should be homogeneous with each other and with the correct answer.

They should fall into the same category as the correct answer (e.g.,: all diagnoses, tests, treatments, prognoses, or disposition alternatives). All distractors should be plausible, grammatically consistent, logically compatible, and relatively the same length as the correct answer.

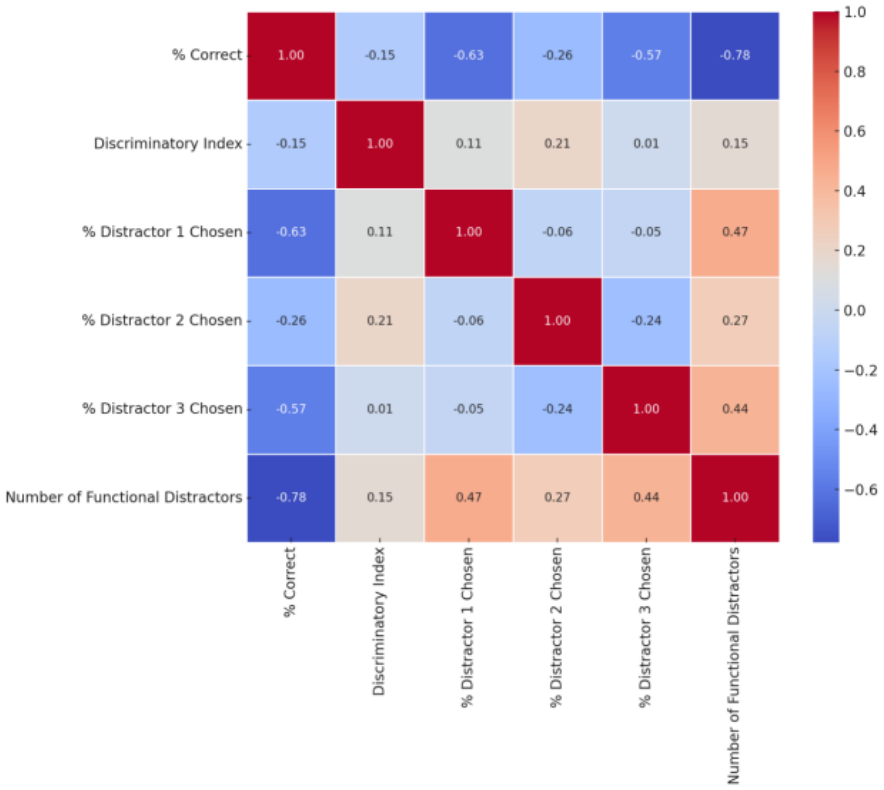
- Each distractor should be plausible and none should stand out as being obviously incorrect. Common misconceptions and faulty reasoning provide a good source of plausible options, as do the mistakes that are often made by a minimally competent candidate. You should be able to

provide the line of reasoning that a candidate would use to select any one of the distractors as an answer. If you are unable to provide the line of reasoning then more than likely, the distractor is not plausible.

- Do not use “All of the above” or “None of the above” as options.

APPENDIX B. Supplementary Figures and Tables

Supplementary Figure 1. Correlation analysis of GPT-4 generated questions**.**



Supplementary Table 1. Summary of six questions requiring revision from initial large language model output	
Questions requiring revising	
Original GPT-4 output	**Revised output (changes in bold)**
Stem: A 4-year-old boy with a history of spina bifida (SB) and associated neurogenic lower urinary tract dysfunction (NLUTD) presents for a routine follow-up. He has been on clean intermittent catheterization (CIC) and antimuscarinic therapy since the age of 3, following the discovery of bladder hostility on urodynamic studies. His recent renal/bladder ultrasound showed stable upper tract findings without any significant hydronephrosis. However, his mother reports increasing episodes of urinary tract infections over the past 4 months. The child's latest serum creatinine was within normal limits. Lead-In: What is the most appropriate next step in the management of this patient?	Stem: A 4-year-old boy with a history of spina bifida (SB) and associated neurogenic lower urinary tract dysfunction (NLUTD) presents for a routine follow-up. He has been on clean intermittent catheterization (CIC) and antimuscarinic therapy since the age of 3, following the discovery of bladder hostility on urodynamic studies. His recent renal/bladder ultrasound showed stable upper tract findings without any significant hydronephrosis. However, his mother reports increasing episodes of urinary tract infections (**that are culture proven and symptomatic**) over the past 4 months. The child's latest serum creatinine was within normal limits. Lead-In: What is the most appropriate next step in the management of this patient
Stem: A 40-year-old woman, with no significant medical history, presents with right flank pain and microscopic hematuria. A CT scan reveals a 10 mm right proximal ureteral stone with moderate hydronephrosis. Her vital signs are stable, and laboratory tests show no signs of infection or renal impairment. Lead-In: What is the most appropriate next step in the management of this patient?	Stem: A 40-year-old woman, with no significant medical history, presents with right flank pain and microscopic hematuria. A CT scan reveals a 10 mm right proximal ureteral stone with moderate hydronephrosis. **The stone appears radio-opaque on scout image.** Her vital signs are stable, and laboratory tests show no signs of infection or renal impairment. Lead-In: What is the most appropriate next step in the management of this patient?

Kim et al. The use of AI (GPT-4) generated multiple-choice questions for examination of surgical subspecialty residents: Report of feasibility and psychometric analysis

<table>
<tr><td>Stem:
A 60-year-old man with a history of diabetes and hypertension presents for a follow-up visit. A year ago, he was found to have bilateral adrenal incidentalomas, each measuring less than 2 cm with benign radiological features. Initial hormonal workup was unremarkable. His diabetes and hypertension are well-controlled with medication.

Lead-In:
What is the most appropriate management for this patient at this follow-up visit?</td><td colspan="2">Stem:
A 60-year-old man with a history of diabetes and hypertension presents for a follow-up visit. A year ago, he was found to have bilateral adrenal incidentalomas, each measuring less than 2 cm with benign radiological features with Hounsfield Units <10. Initial hormonal workup was unremarkable. His diabetes and hypertension are well-controlled with medication.

Lead-In:
What is the most appropriate management for this patient at this follow-up visit?</td></tr>
<tr><td colspan="3">Options requiring revising</td></tr>
<tr><td>MCQ Stem:
A 40-year-old man, involved in an industrial accident, presents with severe pelvic pain, gross hematuria, and difficulty voiding. Examination reveals extensive perineal ecchymosis and a high-riding, tender left testis. A CT scan confirms a pelvic fracture with a bladder neck injury and ultrasound of the scrotum indicates a complex fracture of the left testicular tunica albuginea with extensive parenchymal disruption. The patient is hemodynamically stable but in severe pain.

Lead-In:
Given the complexity of this patient's urological injuries, what is the most</td><td>Options:
A. Immediate cystoscopic evaluation with suprapubic tube placement, followed by delayed scrotal exploration and testicular repair.
B. Immediate scrotal exploration with testicular repair or orchiectomy, and urethral catheterization for bladder drainage.
C. Emergency pelvic external fixation and scrotal exploration.
D. Placement of a urethral catheter, followed by sequential cystoscopic evaluation and scrotal exploration within 24 hours.</td><td>Options:
A. Immediate cystoscopic evaluation with suprapubic tube placement, followed by delayed scrotal exploration and testicular repair.
B. Immediate scrotal exploration with testicular repair or orchiectomy, and urethral catheterization for bladder drainage.
C. Emergency pelvic external fixation, followed by simultaneous open evaluation of bladder injury and scrotal exploration.
D. Placement of a urethral catheter, followed by sequential cystoscopic evaluation and scrotal exploration within 24 hours.</td></tr>
</table>

appropriate initial management strategy?		
Stem: A 65-year-old man presents with back pain and weight loss. His PSA is 30 ng/mL, and a biopsy confirms prostate adenocarcinoma with a Gleason score of 9. Imaging reveals multiple bone metastases, including in the spine, pelvis, and ribs, and a small liver lesion. His performance status is relatively preserved, and he has no significant comorbidities. Lead-In: What is the most appropriate initial therapy for this patient?	Options: A. Androgen deprivation therapy (ADT) alone B. ADT with docetaxel chemotherapy C. Abiraterone acetate and prednisone D. ADT	Options: A. Androgen deprivation therapy (ADT) alone B. ADT with docetaxel chemotherapy C. Abiraterone acetate and prednisone D. ADT with **enzalutamide and docetaxel**
Stem: A 60-year-old woman with a history of osteoporosis and recent onset hypertension is referred for an evaluation of an incidental adrenal mass. The mass was discovered during an abdominal MRI performed for unrelated chronic back pain. The MRI report describes a 4 cm left adrenal mass with heterogeneous enhancement and a slightly irregular border. The mass does not exhibit characteristic features of a benign lipid-rich adenoma. The patient denies any significant weight gain, facial rounding, or striae. Her blood pressure	Options: A. Immediate referral for surgical resection of the adrenal mass. B. Perform a 1 mg overnight dexamethasone suppression test. C. Schedule a biopsy of the adrenal mass to assess for malignancy. D. Initiate a contrast-enhanced washout CT scan to further characterize the adrenal mass	Options: A. Immediate referral for surgical resection of the adrenal mass. B. Perform a 1 mg overnight dexamethasone suppression test. C. Schedule a biopsy of the adrenal mass to assess for malignancy. D. Initiate a contrast-enhanced washout CT scan to further characterize the adrenal mass, **followed by hormonal work up.**

is currently well-controlled with a single antihypertensive agent. She has no history of diabetes or notable family history of endocrine disorders. Lead-In: Considering the patient's clinical presentation and imaging findings, what is the most appropriate initial step in the management of this patient?		

Supplementary Table 2. Per topic psychometric analysis

Topic	Median correct (IQR)	Mean correct	STD correct	Poor DI, n (%)	Acceptable DI, n (%)	Excellent DI, n (%)	Median number of functional distractors (IQR)
Pediatric urology	86.70% (60–100%)	80.14%	21.15%	3 (60%)	2 (40%)	0 (0%)	0.5 (0–2)
Neoplasm	66.70% (46.7–86.7%)	71.18%	23.20%	8 (53%)	5 (33%)	1 (7%)	1.5 (1–2)
Trauma and fistula	66.70% (53.3–100%)	71.67%	18.93%	4 (57%)	2 (29%)	1 (14%)	2 (0–2)
Urinary tract obstruction	53.30% (46.7– 80%)	60.67%	18.57%	2 (40%)	3 (60%)	0 (0%)	1 (1–2)
Urolithiasis	53.30% (46.7– 73.3%)	61.87%	18.29%	1 (14%)	3 (43%)	3 (43%)	1 (1–2)

Neurourology and voiding dysfunction	33.30% (20–46.7%)	33.02%	12.71%	1 (14%)	3 (43%)	3 (43%)	2 (1–2)
Andrology and endocrinology	60% (33.3–66.7%)	56.27%	25.22%	7 (47%)	3 (20%)	5 (33%)	2 (1–2)
Infection and inflammation	46.70% (20–66.7%)	49.01%	27.16%	4 (50%)	3 (37%)	1 (13%)	2 (1–2)
Transplantation, nephrology, renovascular disease	53.30% N/A	53.30%	N/A	0 (0%)	1 (100%)	0 (0%)	2 (NA)
Diagnostic techniques and imaging	53.30% (13.3–100%)	64.50%	35.31%	2 (50%)	1 (25%)	1 (25%)	1 (0–2)
Patient safety	0% NA	0.00%	N/A	1 (100%)	0 (0%)	0 (0%)	3 (NA)
Basic sciences	76.70% (53.3–100%)	76.65%	23.47%	1 (50%)	1 (50%)	0 (0%)	1 (0–2)