Human-in-the-loop AI: Enhancing Assessment Item Generation in Medical Education

Xiaomei Song, PhD, Case Western Reserve University

Kate Weber, PhD, Case Western Reserve University

A. A list of Gen AI tools

ChatGPT	• Generates MCQs, short answer, essay, matching, and case-based questions • Strong for clinical vignettes and reasoning-based questions • Highly customizable with prompts (e.g., OSCE, EOR, NBME-style items) • ChatGPT 4.0 fast for bulk question generation
Gemini	• Structured question generation from documents (PDFs, notes, slides) • Strong for summarization → question conversion workflows • Integrates with Google ecosystem (Docs, Drive, Classroom)
Claude	• High-quality long-form question writing • Strong at clinical reasoning and nuanced MCQ stems • Handles long documents very well (large context window) • Often preferred for assessment quality and clarity of wording
Copilot	• Generates questions from Word, PowerPoint, and web content • Strong integration with Microsoft 365 (Teams, Forms, Word) • Useful for educators already in the Microsoft ecosystem
NotebookLM	• Modifies uploaded materials into summaries, flashcards, and practice questions • Very strong for course-based exam preparation • Ideal for converting lecture notes into question banks
Custom GPTs / AI Builders (e.g., built on ChatGPT, Claude, Gemini, etc.)	• Modifies uploaded materials into practice questions for formative purposes • Very strong for course-based question preparation • Ideal for converting lecture notes and materials into question banks

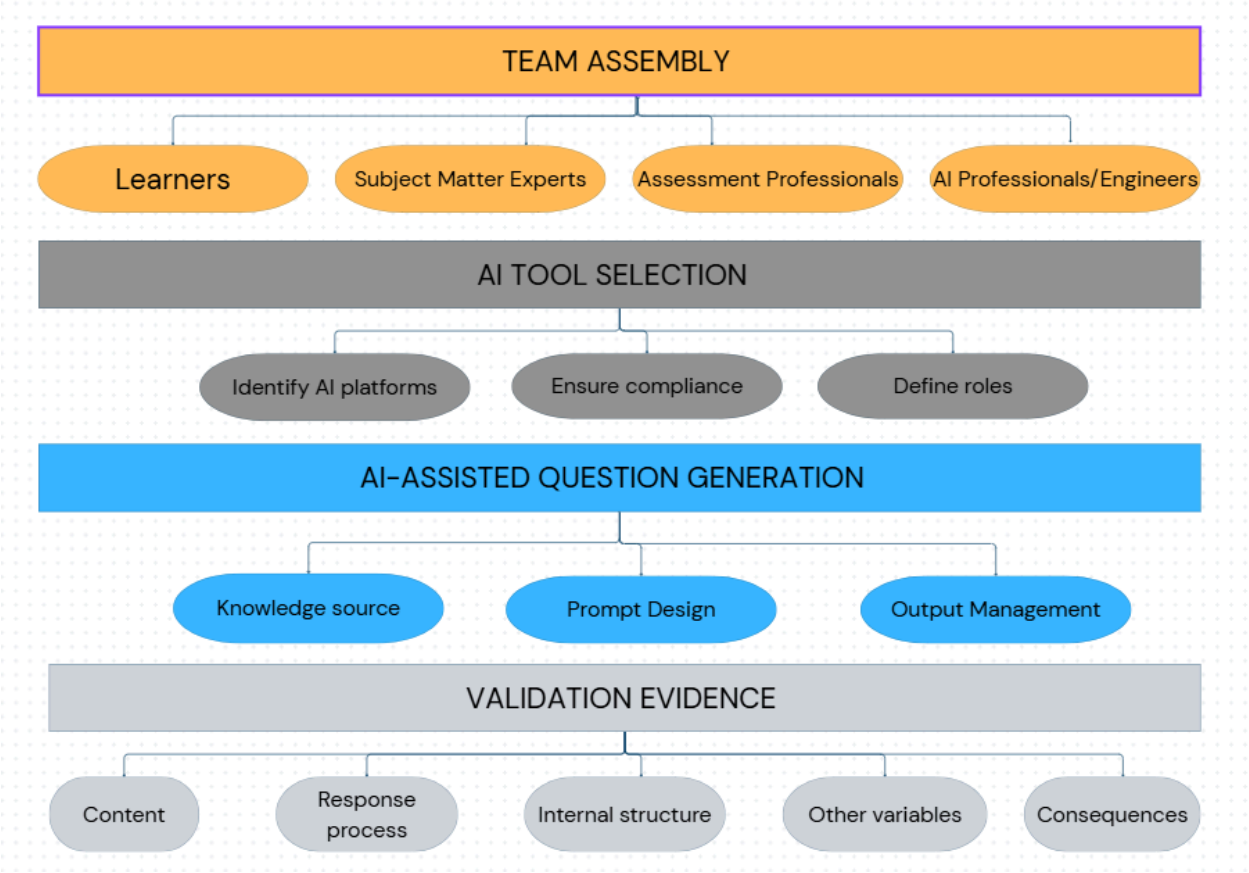
B. Draft Prompt

Text	Comment
Context: You are an expert in evaluation methods preparing exam questions about Grant Writing.	Context helps the model frame its response in the terms you want. "You are a curious ten-year-old" would get completely different output
Using only the provided resources, generate one USMLE-style multiple-choice question for each learning objective in the 'Grant Writing LOs' document.	You are also telling it what information to use and giving it the overall task.
Question Structure: **Clinical Vignette:** Provide a realistic scenario. **Lead-in:** A clear, single question. **Options:** Exactly five (A-E). **Key:** The correct answer.	You explicitly define both the content and the format of the response.
Explanation: For each question, provide a 'Rationale' for the correct answer and a 'Distractor Analysis' explaining why each incorrect option is plausible but wrong. Identify the linked Learning Outcome.	The response must include justification from the model for its decisions to help the human reviewer assess the question.
Constraints: Focus on application and analysis (Bloom's Taxonomy level 3+). Avoid 'all of the above' or 'none of the above.' Ensure the correct answer letter is randomized across the set.	Key constraints to shape how the questions are prepared

Library:

- https://gail.wharton.upenn.edu/prompt-library/
- https://mitsloanedtech.mit.edu/ai/basics/effective-prompts/
- https://structuredprompt.com/wp-content/uploads/2023/08/TRACI-Users-Guide-v1.pdf

C. Human-in-the-loop AI Item Generation Flowchart and Checklist



Team Assembly	
☐	Learners
☐	Subject matter experts
☐	Assessment professionals
☐	AI professionals/engineers
Tool Design	
☐	Identify AI platforms capable of question/item generation
☐	Ensure data privacy, security, and institutional compliance
☐	Define roles for AI vs. human review in workflow
AI-assisted Question Generation	
☐	Knowledge source

☐	Prompt design
☐	Output management
Validation Evidence	
Content validity (see examples)	
☐	Alignment: Ensure AI-assisted questions align with stated learning objectives and instructional content
☐	Source Mapping/Representativeness: Confirm that each AI-assisted question clearly maps to a specific page or section in the provided material
☐	Accuracy: Verify that each AI-assisted question is based on correct, evidence-based information (i.e., free from hallucinations or inaccuracies)
☐	Difficulty Balance: Include a mix of both challenging and easy AI-assisted questions
☐	Clinical Vignettes: Clinical scenarios may extend beyond the immediate topic; it is acceptable if questions can be answered without the vignette, as broader exposure supports learning
☐	Distractor Quality: Ensure all AI-assisted question answer choices (distractors) are plausible and functionally relevant
☐	Bias Review: Avoid using demographic characteristics (e.g., age, gender, race) as diagnostic shortcuts unless clinically justified
☐	Images & Media: When an AI-assisted question references an image, ensure an appropriate, high-quality image is selected and inserted based on the description
☐	Audit Trail: Document reviewer feedback, approvals, and edits for transparency and tracking
Response process (see examples)	
☐	Think-Aloud Protocols: Ask students to verbalize their thought process while answering AI-assisted questions to identify misinterpretation or unintended cues
☐	Cognitive Interviews: Conduct cognitive interviews with learners to confirm that their reasoning processes align with the intended difficulty and cognitive level
☐	Reasoning Alignment: Confirm with students that correct responses require the targeted reasoning (e.g., application, analysis) rather than test-taking strategies or guesswork
Internal structure (see examples)	
☐	Item Difficulty: Analyze item difficulty (e.g., p-values) to ensure an appropriate range across AI-assisted questions
☐	Item Discrimination: Evaluate how well AI-assisted items differentiate between high- and low-performing learners (e.g., discrimination index, point-biserial correlation)
☐	Reliability: Assess overall test reliability (e.g., Cronbach's alpha or KR-20 for dichotomous items)
☐	Dimensionality: Evaluate whether AI-assisted items measure the intended construct(s) (e.g., factor analysis)
☐	Score Consistency: Check for consistency of scores across years
Relationships with other variables (see examples)	
☐	Criterion-Related Validity: Evaluate how well scores predict relevant outcomes (e.g., clinical performance, course success)

☐	Convergent Evidence: Determine whether scores correlate with other measures of the same or similar constructs (e.g., course exams, standardized tests)
☐	Discriminant Evidence: Confirm low correlation with unrelated constructs to ensure the test is not measuring unintended skills
☐	External Benchmarks: Compare performance to established standards or norms when available
Consequences (see examples)	
☐	Intended/unintended Impact: Evaluate whether the assessment supports or hinders learning, retention, teaching and desired educational outcomes
☐	Fairness: Ensure the assessment does not advantage or disadvantage specific groups and supports equitable evaluation
☐	Decision Accuracy: Review whether decisions based on scores (e.g., pass/fail, remediation) are appropriate and justified

Verbatim Suggested Prompt

You are an expert in assessment of medical students and development of multiple-choice exams. Using only the provided resources, generate one USMLE-style multiple-choice question for each learning objective in the 'Grant Writing LOs' document.

Structure:

Clinical Vignette: Provide a realistic scenario.

Stem: A clear, single question.

Options: Exactly five (A-E).

Key: The correct answer.

Explanation: For each question, provide a 'Rationale' for the correct answer and a 'Distractor Analysis' explaining why each incorrect option is plausible but wrong. Identify the linked Learning Outcome.

Constraints: Focus on application and analysis (Bloom’s Taxonomy level 3+). Avoid 'all of the above' or 'none of the above.' Ensure the correct answer letter is randomized across the set.