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

Xiaomei Song, PhD, Director of Student Assessment, Associate Professor
Kate Weber, PhD, Director of AI in Medical Education, Assistant Professor

CASE WESTERN RESERVE UNIVERSITY
School of Medicine

Objectives

By the end of this session, participants will be able to

1. Understand the concepts of and explain applications of Generative AI / LLMs in assessments
2. Summarize foundational capabilities and limitations of Gen AI
3. Apply human-in-the-loop strategies grounded in validity theory to AI-based item generation
4. Have hands-on experience with free AI tools

Handout!

CASE WESTERN RESERVE UNIVERSITY
School of Medicine

Assessment in Medical Education

Assessments at our school

Diverse assessment formats

- MCQs, short-answer, essays, practical exams

Diverse purposes

- TBL sessions, reflections, assignments, summative exams

High-quality assessment is essential for

- **Valid measurement** of knowledge and skills
- **Fair and defensible** decisions

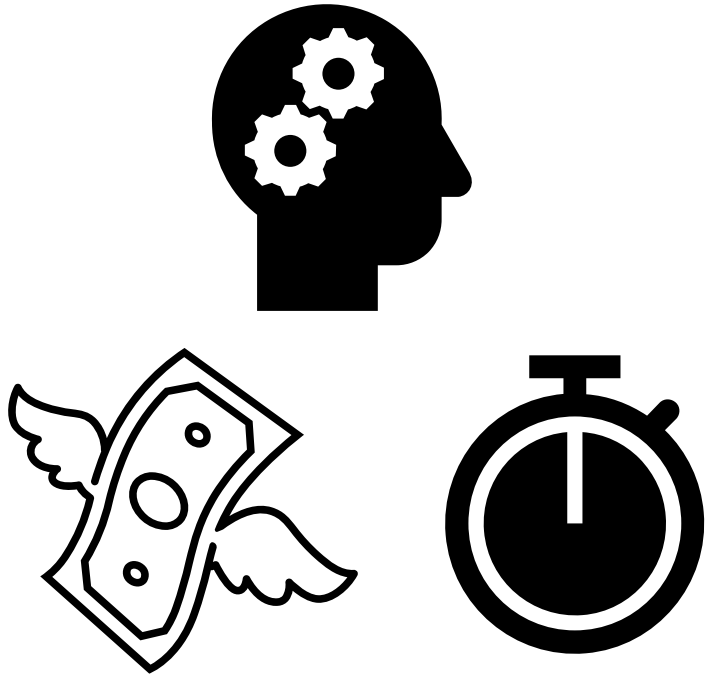
Challenges in Assessment Development

Developing quality items is

- Time-intensive
- Costly
- Expertise-dependent

Common institutional challenges

- Limited faculty time
- Variable assessment training
- Budget constraints
- Need for curriculum-specific alignment



CASE WESTERN RESERVE UNIVERSITY
School of Medicine

Automatic Item Generation (AIG)

Automatic Item Generation (AIG)

- Uses cognitive models and item templates
- Produces multiple assessment items efficiently

Benefits

- Scalability
- Consistency
- Reduced development time

Historically **limited by**

- Technical complexity
- Rigid templates

Gierl M, Swygert K, Matovinovic D, Kulesher A, Lai H. Three Sources of Validation Evidence Needed to Evaluate the Quality of Generated Test Items for Medical Licensure. Teaching and Learning in Medicine. 1-11. PMID 36106359 DOI: 10.1080/10401334.2022.2119569

CASE WESTERN RESERVE UNIVERSITY
School of Medicine

Opportunity: Generative AI & LLMs

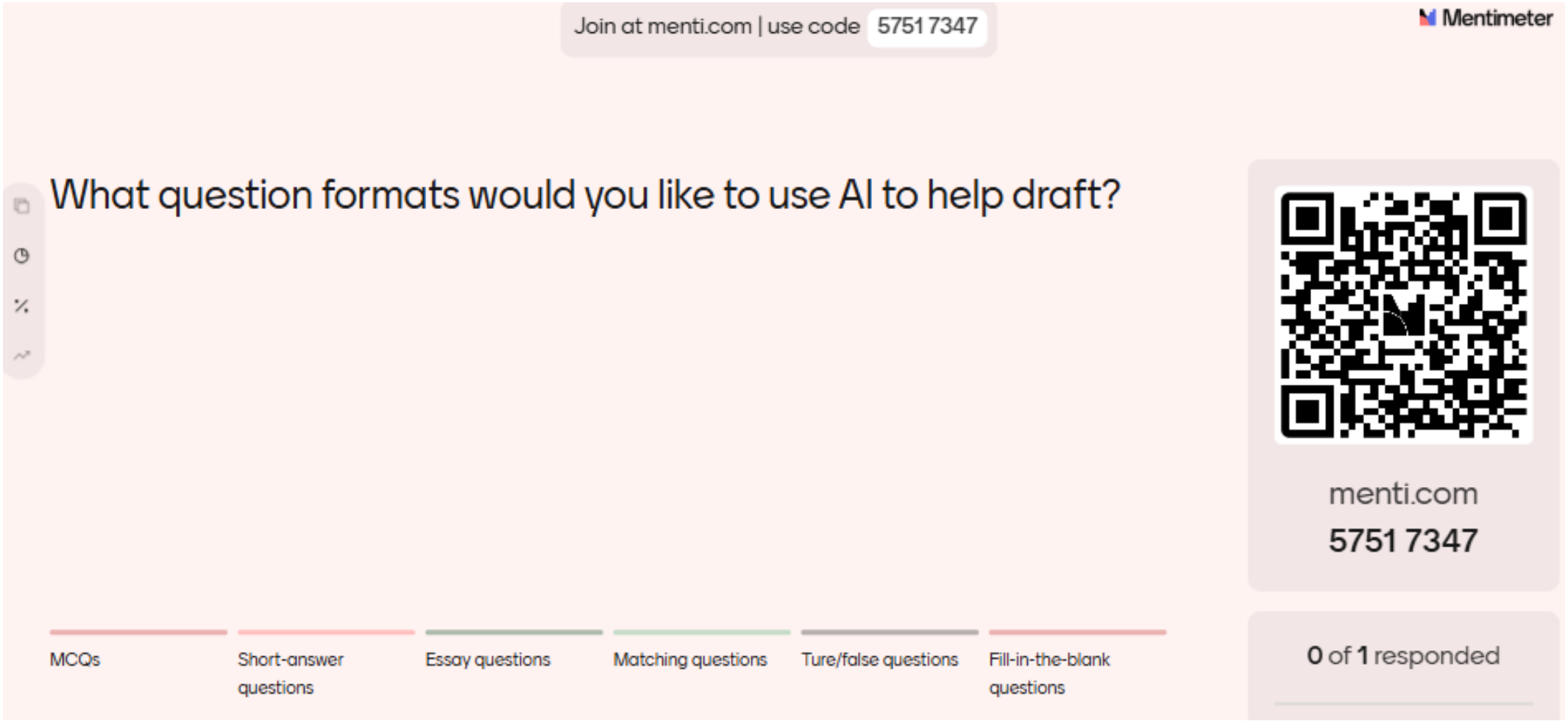
New capabilities

- Natural language generation
- Context-aware item creation
- Rapid iteration

Growing institutional interest

- Educational technology
- AI-supported teaching and assessment

Poll: Assessment Tools



https://www.mentimeter.com/app/presentation/al3v7nh1twxom1jcafh2t9b6737orspc/present?question=2q33p7sv7wdr

Questions or comments?

CASE WESTERN RESERVE
UNIVERSITY
School of Medicine

Gen AI Capacities and Limitations

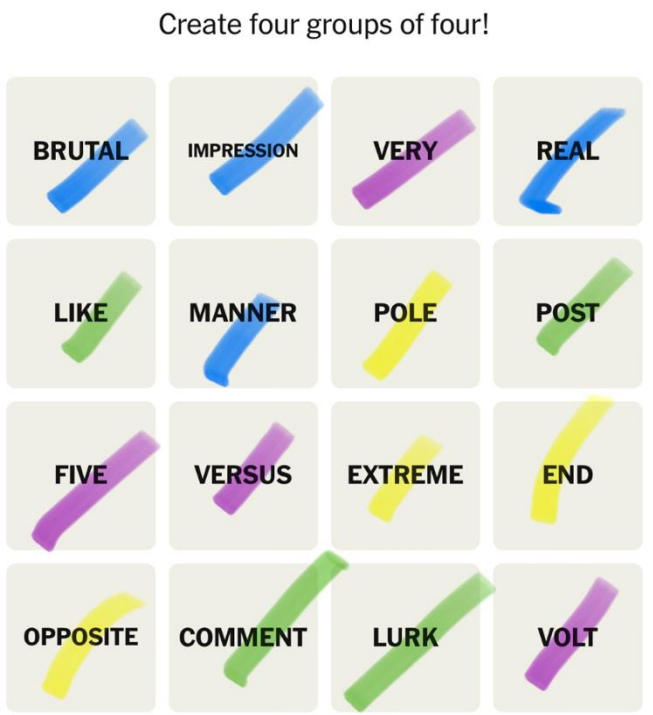
- AI systems trained on vast amounts of text
- Learn patterns in language
- Generate human-like text by predicting what comes next
- NOT search engines - they're prediction machines
- Must look up real-time information or request internet access (without tools)

Gen AI: What they CAN do

- Writing assistance (drafting, editing, summarizing)
- Research support (literature review, synthesis)
- Data analysis and interpretation
- Brainstorming and ideation
- Learning and explanation
- Translation and transformation
- Programming

Gen AI: Where they Stumble

- **No real-time data** (unless specifically connected)
- **Hallucinations** Can confidently state false information
- **No reasoning in traditional sense** pattern matching
- **Context limits** Can't remember everything forever
- **Biases** Reflect biases in training data
- **Math/counting** Often struggle with arithmetic
- **Medical Image Generation** Not ready for prime-time

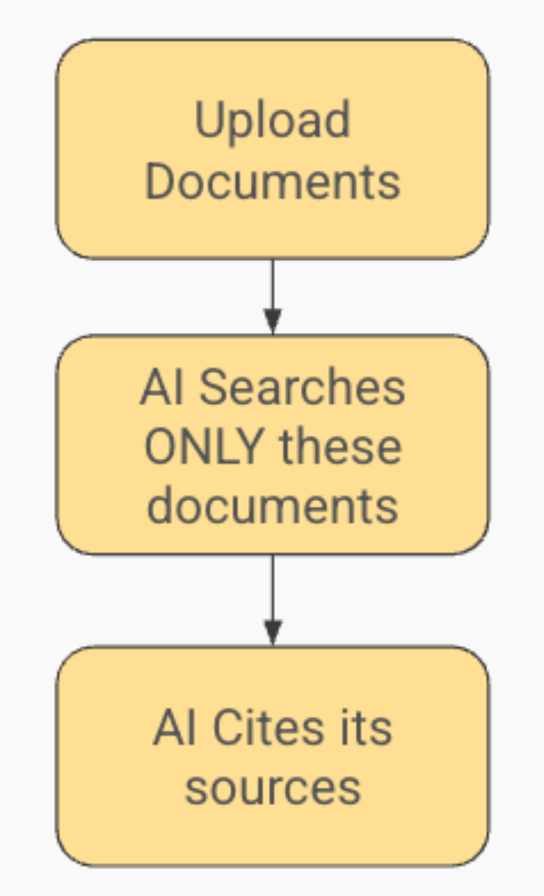


CASE WESTERN RESERVE UNIVERSITY
School of Medicine

Grounded AI: Provide it with YOUR sources

Retrieval **A**ugmented **G**eneration (**RAG**)

- **Retrieval** finds relevant parts of your documents
- **Augmented** Adds that information to the prompt
- **Generation** Creates an answer based on that information



CASE WESTERN RESERVE UNIVERSITY
School of Medicine

Grounded AI: Live Example

NotebookLM

- Free* tool from Google
- Ingests knowledge sources
- Cites sources
- Synthesizes Information

https://notebooklm.google.com/notebook/09f383dc-c84a-4d0e-a4b9-027de5b4e6a0

CASE WESTERN RESERVE UNIVERSITY
School of Medicine

Building AI-Assisted Item Generation

- Find your **Team**
- **Design** the Generation System
- **Generate** Candidate Questions
- **Validate** the Results

CASE WESTERN RESERVE
UNIVERSITY
School of Medicine

The Team

- **Learners**
- **Subject-Matter Experts**
- **Assessment Professionals**
- **AI Professionals / Engineers**

CASE WESTERN RESERVE UNIVERSITY
School of Medicine

Designing AI-assisted Item Generation

- **Foundation Model / Platform** (e.g., Gemini, NotebookLM, Custom)
- **Knowledge Sources** (e.g., curriculum, reference material, papers)
- **Prompt Design** shapes the system
  - Define **population and cognitive level**
  - Control **style and difficulty**
  - **Output constraints** (e.g., format, level, objectives)

The Prompt

Context

Using only the provided resources, generate one USMLE-style multiple-choice question for each learning objective in the 'Grant Writing LOs' document.

Expectations

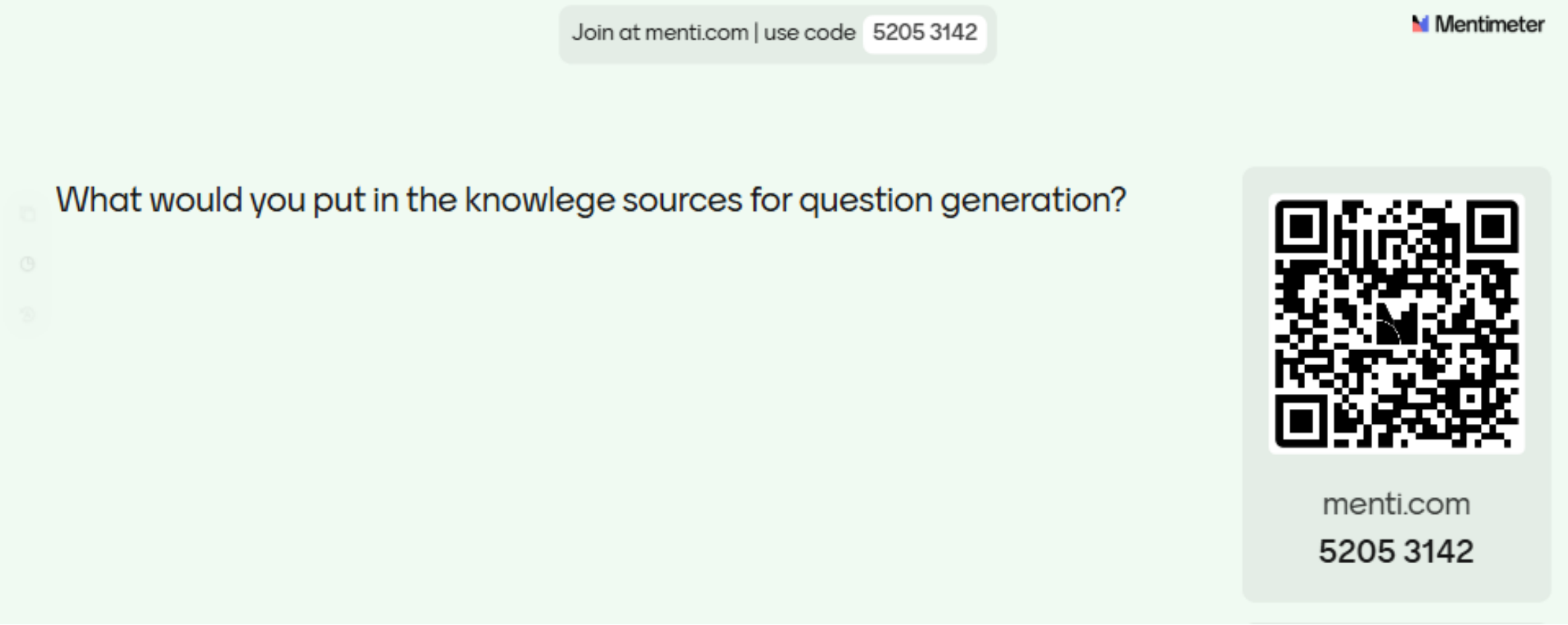
Structure:

- **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

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.

Poll: Knowledge sources



https://www.mentimeter.com/app/presentation/al9dyp1v7gq79vacgn8uxy563qwt6rz3/edit?question=xa421efs5wg9

Questions or comments?

CASE WESTERN RESERVE
UNIVERSITY
School of Medicine

Why Human-in-the-Loop?

Human-in-the-Loop

- Intentional inclusion of human judgment
- Before, during, or after AI output

Essential for

- Validity
- Fairness
- Ethical use

Balances

- Efficiency of AI
- Expertise of educators

Quality Control Matters

AI-generated items must meet standards for

- Assessment validity
- Alignment with learning objectives and instruction
- Fairness and inclusivity

Risks and challenges

- Biases and misalignment
- Ethical concerns, transparency, authorship, and trust
- AI uses and acceptance

HITL Design: Content validity

Definition

Degree to which items reflect intended learning objectives and curriculum

Key questions, e.g.

Is this item relevant or representative to what learners are expected to know or do?

Human Stakeholders

- Subject matter experts (SMEs), AI professionals/engineers, assessment professionals

HITL checkpoints

- Review AI prompts for objective alignment
- Review knowledge sources,
- Review stems and options for ambiguity
- Post-generation screening for relevance and accuracy

HITL Design: Response Process

Definition

Evidence that items elicit intended reasoning or decision-making processes

Key questions, e.g.

Are learners thinking in the way we intend?

Human Stakeholders

- Learners, SMEs, assessment professionals

HITL checkpoints

- Analyze whether reasoning aligns with instructional intent (e.g., "I don't need to read stems")

HITL Design: Internal Structure

Definition

Relationships among items and consistency with assessment design

Key questions, e.g.

- *Do these items function together as a coherent assessment?*

Human Stakeholders

- Assessment professionals, SMEs

HITL checkpoints

- Compare AI-generated items within and across assessments
- Monitor post-administration performance trends

HITL Design: Relations to other variables

Definition

Evidence that scores relate appropriately to other measures (e.g., performance, level of training)

Key questions, e.g.

Do student performance align with what we expect in other assessments based on learner characteristics?

Human Stakeholders

- Assessment professionals, SMEs

HITL checkpoints

- Review feedback and learner experiences
- Reassess items used for high-stakes decisions

HITL Design: Consequences

Definition

Impact of assessment use on learners, faculty, and programs

Key questions, e.g.

What are the educational and ethical consequences of using these AI-generated items?

Human Stakeholders

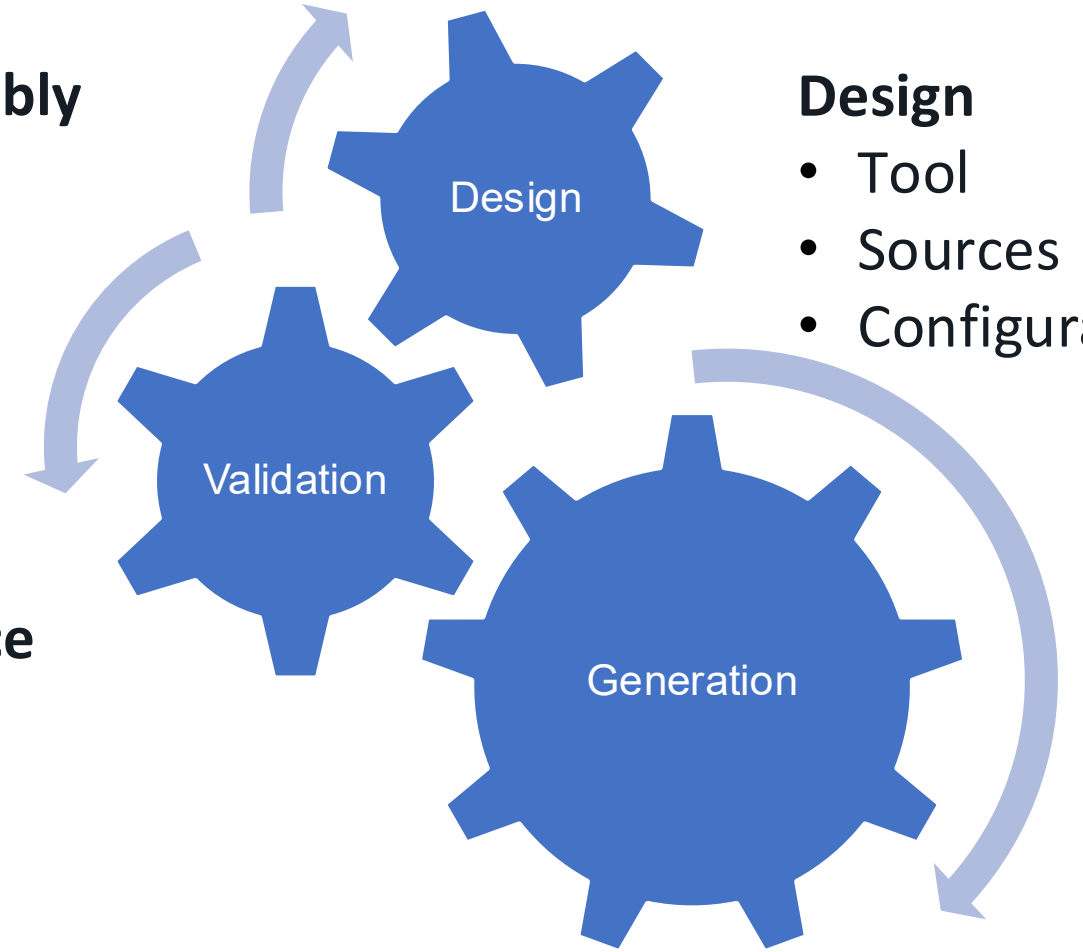
- Learners, SMEs, assessment professionals, the whole community

HITL checkpoints

- Investigate inferential statistics with other assessment results

HITL Item Generation

Team Assembly



Design

- Tool
- Sources
- Configuration

Initial Question Generation Core Components

- Knowledge source
- Prompt design
- Output Management

Validity Evidence Accumulation

Questions or comments?

CASE WESTERN RESERVE
UNIVERSITY
School of Medicine

HITL Item Generation Hands-on Activity

YOUR TURN

HITL Item Generation Hands-on Activity

Option A: Imagine you have a team and plan to use NotebookLM to generate questions. How would you approach it? Use the checklist and work together to:

1) identify the knowledge sources
2) draft the prompts
3) generate 1–3 questions
4) suggest approaches to collect validity evidence

HITL Item Generation Hands-on Activity

Option B: The knowledge sources have already been identified and questions related to **principles of grant writing** have been generated using Notebook LM.

Use the checklist and work together to:

1. Write a question-generating prompt
2. review the quality of the questions
3. suggest approaches to collect validity evidence

Group debrief

- Key observations
- Common issue identified
- Effective strategies for improvement

Key Takeaways

- Clearly define your goals, tools, and overall process.
- Keep humans meaningfully involved—human-in-the-loop design is essential.
- Move beyond generating questions to thoughtfully validating them.
- Maintain intentional oversight to ensure quality, validity, and ethical integrity.

Contact Information

Xiaomei.Song@case.edu
kate-weber@case.edu

CASE WESTERN RESERVE
UNIVERSITY
School of Medicine