Using AI to Create, Assess & Align Multiple Choice Questions

Can AI help align assessments to Bloom's Taxonomy?

Karen Chen (Assistant Professor, IS) & Fow-Sen Choa (Professor, CSEE)
Kevin Hwang & Kenneth Wang (Glenelg High School)
Maryam Alomair (IS PhD student)

This presentation is based on paper presented at AI in Education Conference 2024

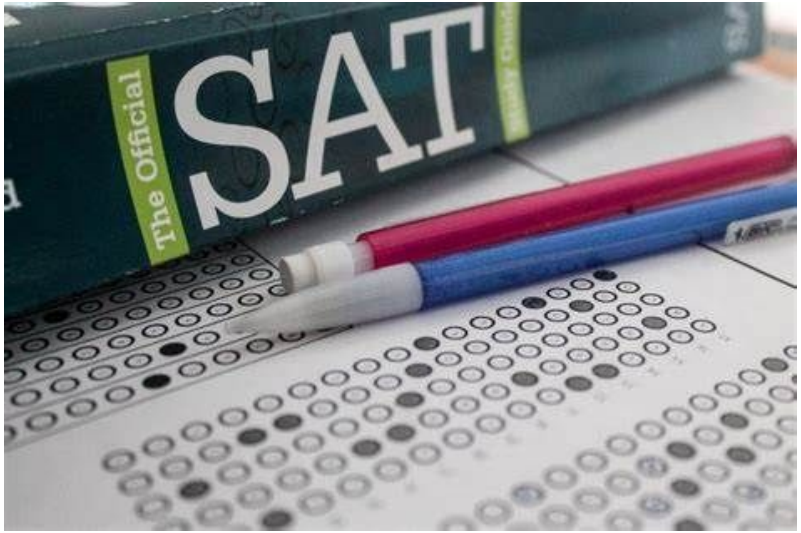
Hwang, Kevin, Kenneth Wang, Maryam Alomair, Fow-Sen Choa, and Lujie Karen Chen. "Towards Automated Multiple Choice Question Generation and Evaluation: Aligning with Bloom's Taxonomy." In *International Conference on Artificial Intelligence in Education*, pp. 389-396. Cham: Springer Nature Switzerland, 2024.

Special Acknowledgement

Kevin Hwang & Kenneth Wang

- High school seniors from Glenelg High School
- For their excellent contribution to AI in Education research to help to improve the quality of Multiple Choice Questions -:)
- BTW: How did you get interested in this project?

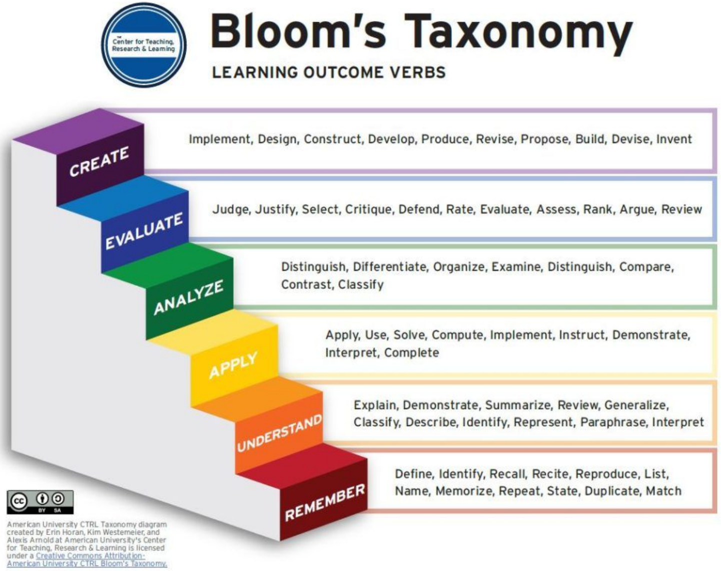
Multiple Choice Questions (MCQs): Easy to Grade, Difficult to Generate



SAT: standardized US exam, including 154 MCQs

Image credits: https://testprephq.com/about-the-sat/

Well-designed
Well-aligned



Bloom's Taxonomy

LEARNING OUTCOME VERBS

CREATE	Implement, Design, Construct, Develop, Produce, Revise, Propose, Build, Devise, Invent
EVALUATE	Judge, Justify, Select, Critique, Defend, Rate, Evaluate, Assess, Rank, Argue, Review
ANALYZE	Distinguish, Differentiate, Organize, Examine, Distinguish, Compare, Contrast, Classify
APPLY	Apply, Use, Solve, Compute, Implement, Instruct, Demonstrate, Interpret, Complete
UNDERSTAND	Explain, Demonstrate, Summarize, Review, Generalize, Classify, Describe, Identify, Represent, Paraphrase, Interpret
REMEMBER	Define, Identify, Recall, Recite, Reproduce, List, Name, Memorize, Repeat, State, Duplicate, Match

American University CTRL Taxonomy diagram created by Erin Horan, Kim Westemeier, and Alexis Arnold at American University's Center for Teaching, Research & Learning is licensed under a Creative Commons Attribution-American University CTRL Bloom's Taxonomy.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. Theory into practice, 41(4), 212-218.

Can LLM (e.g. GPT 4) be used to generate **high quality** Multiple Choice Questions (MCQs) that are **aligned with Bloom's Taxonomy**?

Emerging Research in Bloom's Taxonomy-aligned Question Generation

- Kusuma et al. 2018
  - utilized templates with pattern matching and keyword identification
  - did not validate the generated Bloom's taxonomy level
- Elkins et al. 2023
  - utilized InstructGPT to generate free response questions aligned to Bloom's taxonomy

Kusuma, Selvia Ferdiana, and Rinanza Zulmy Alhamri. Generating Indonesian question automatically based on Bloom's taxonomy using template based method. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control (2018): 145-152.

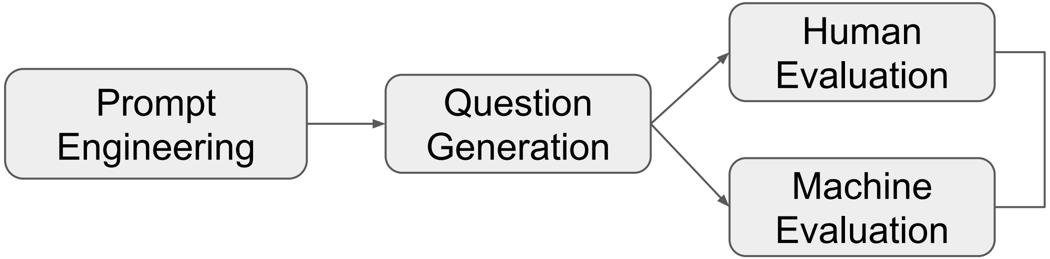
Elkins, Sabina, et al. How Useful are Educational Questions Generated by Large Language Models?. International Conference on Artificial Intelligence in Education. Cham: Springer Nature Switzerland, 2023.

Emerging Research in Bloom's Taxonomy-aligned Question Generation (cont'd)

- Scaria et al. (AIED 2024)
    - Used few-shot and Chain of Thoughts prompting on five different LLMs
    - Show that LLMs have the capability of generating high quality & diverse free-response questions that have accurate alignment to Bloom's taxonomy.

Scaria, N., Dharani Chenna, S., & Subramani, D. (2024, July). Automated Educational Question Generation at Different Bloom's Skill Levels Using Large Language Models: Strategies and Evaluation. In *International Conference on Artificial Intelligence in Education* (pp. 165-179). Cham: Springer Nature Switzerland.

Today's Plan



Research Questions

- How well does GPT-4 generate **usable and relevant** MCQs aligned to Bloom's Taxonomy, according to human evaluation?
- How does GPT-4's performance vary with the number of shots/examples?
- To what extent does the automatic MCQ quality evaluation (e.g. IWF) align with human feedback?

Practical Implications/Discussion

- Instructors' perspectives
- Students' perspectives

Example Prompt with One Shot/Example

Instruction

You are a teacher trying to generate **multiple choice questions** for an upcoming assessment.

You will be given an **excerpt to do with a certain science related topic.**
You will also be given a desired **Bloom's taxonomy level.**
Your task is to generate one multiple choice question, with one correct answer and three distractors each, that correspond with the desired Bloom's taxonomy level. To aid you, I have provided some examples:

One Shot / Example

context: A four important lymphocyte is the natural killer cell, a participant in the innate immune response. A natural killer cell (NK) is a circulating blood [...]
taxonomy: Understand
question: Which of the following cells would be most active in early, antiviral immune responses the first time one is exposed to pathogen?,
answer: Natural Killer Cell
distractor_1: Macrophage
distractor_2: T Cell
distractor_3: Neutrophil

Specific Request

Taxonomy: Understand
Context: Viruses are acellular entities that can usually only be seen with an electron microscope. Their genomes contain either DNA or RNA, and they replicate using the replication proteins of a host cell. Viruses are diverse [...]

GPT Output

question: In the context of viral infections what is meant by a 'productive' viral infection?,
answer: A productive viral infections results in new virions being produced within the host cell.
distractor_1: A productive viral infection refers to the virus remaining inside the cell without producing new virions.
distractor_2: A productive viral infection denotes the complete destruction of the host cell.
distractor_3: A productive viral infection stimulates the immune system to destroy the virus without the production of new virions.

Example Prompt with Multiple Shots/Examples

Instruction

You are a teacher trying to generate **multiple choice questions** for an upcoming assessment.

You will be given an **excerpt to do with a certain science related topic.**
You will also be given a desired **Bloom's taxonomy level.**
Your task is to generate one multiple choice question, with one correct answer and three distractors each, that correspond with the desired Bloom's taxonomy level. To aid you, I have provided some examples:

Multiple Shots / Examples

context: A four important lymphocyte is the natural killer cell, a participant in the innate immune response. A natural killer cell (NK) is a circulating blood [...]
taxonomy: Understand
question: Which of the following cells would be most active in early, antiviral immune responses the first time one is exposed to pathogen?,
answer: Natural Killer Cell
distractor_1: Macrophage
distractor_2: T Cell
distractor_3: Neutrophil

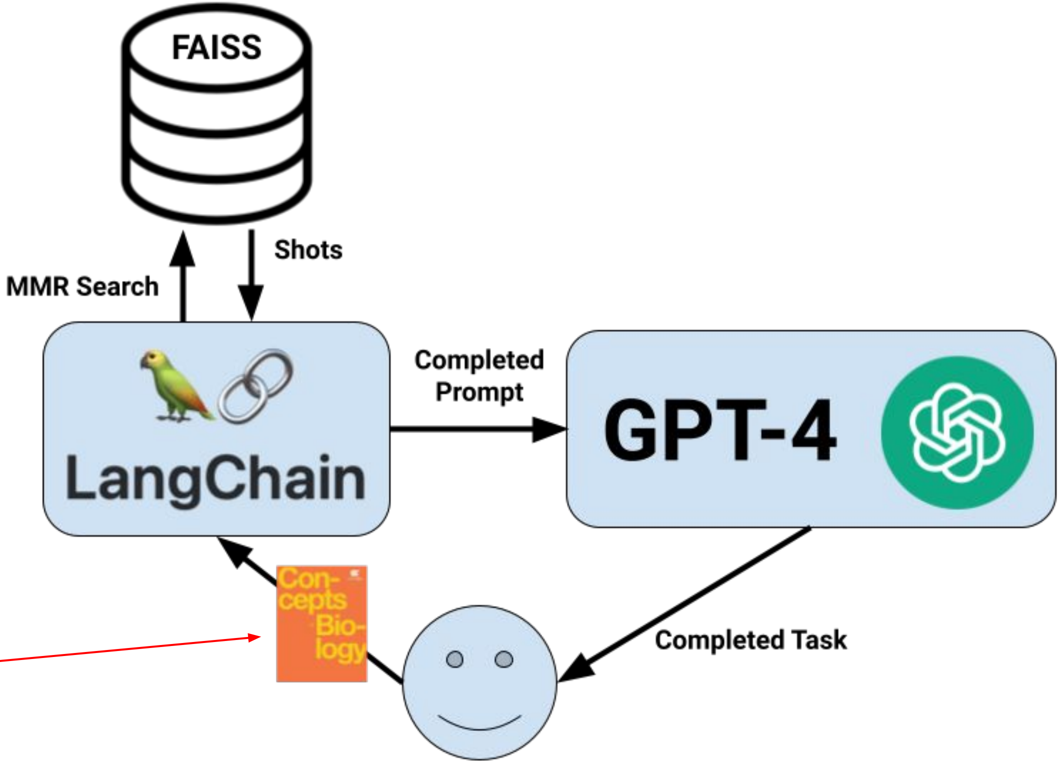
Specific Request

Taxonomy: Understand
Context: Viruses are acellular entities that can usually only be seen with an electron microscope. Their genomes contain either DNA or RNA, and they replicate using the replication proteins of a host cell. Viruses are diverse [...]

GPT Output

question: In the context of viral infections what is meant by a ‘productive’ viral infection?,
answer: A productive viral infections results in new virions being produced within the host cell.
distractor_1: A productive viral infection refers to the virus remaining inside the cell without producing new virions.
distractor_2: A productive viral infection denotes the complete destruction of the host cell.
distractor_3: A productive viral infection stimulates the immune system to destroy the virus without the production of new virions.

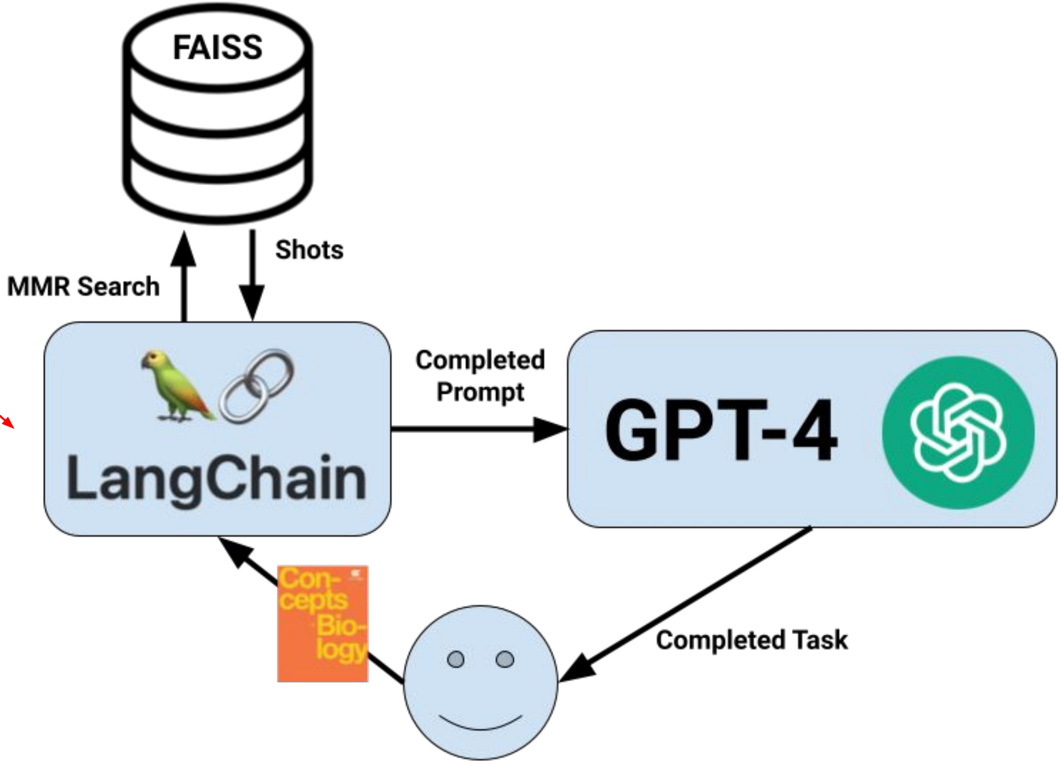
Question Generation



Provide relevant biology textbook excerpt

Question Generation

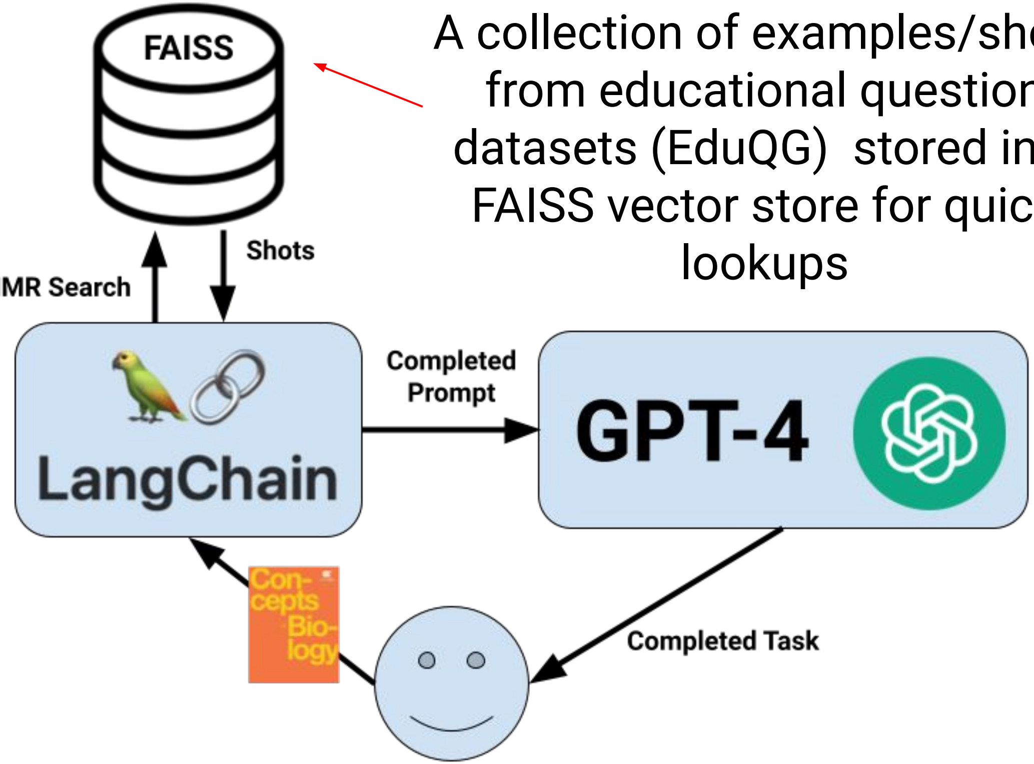
Use LangChain to streamline the workflow (e.g. adds formatting instructions for GPT-4 and generates the GPT-4 prompt)



Question Generation

To select shots matching the topic

Shots retrieved via Maximal Marginal Relevance **(Carbonell et al.)** search to optimize for diversity and relevance.

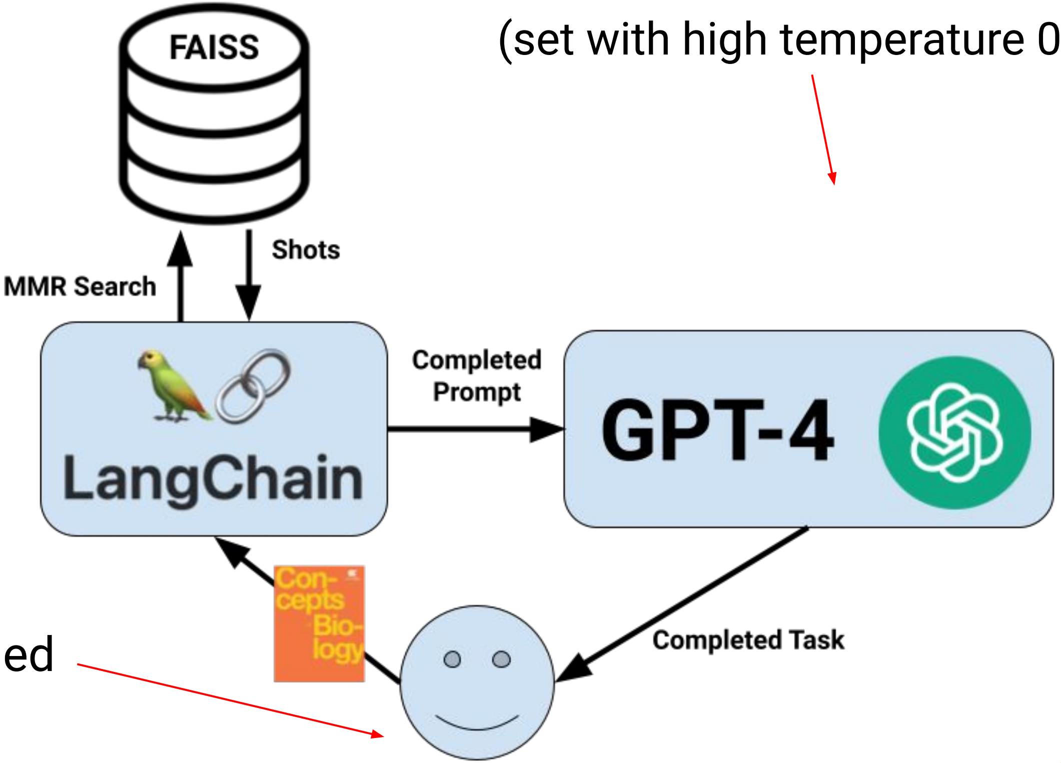


A collection of examples/shots from educational question datasets (EduQG) stored in a FAISS vector store for quick lookups

Carbonell, J., & Goldstein, J. (1998, August). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 335-336).

Question Generation

Prompt with relevant excerpt
and shots fed to GPT-4
(set with high temperature 0.9)



Set of MCQs returned

Question Generation Experiment Matrix (Shot = Example)

	0 Shot	1 shot	3 shot	5 shot	7 shot	Total
Remember	10	10	10	10	10	**50**
Understand	10	10	10	10	10	**50**
Apply	10	10	10	10	10	**50**
Analyze	10	10	10	10	10	**50**
Evaluate	10	10	10	10	10	**50**
Total	**50**	**50**	**50**	**50**	**50**	**250**

Question Evaluation

Human evaluation

- Two human experts evaluated a random sample (n=150 stratified sample out of 250 MCQs) of the generated questions, a third expert was used to reconcile differences in evaluation.
- Golden standard, ground truth values to compare against automated evaluation
- Reviewed on **relevance and educational usability** at scales of 1-4
- Reviewed on **Bloom's taxonomy level** (blind from GPT output)

Moore, S., Nguyen, H. A., Chen, T.., Stamper, J., Assessing the quality of multiple-choice questions using GPT-4 and rule-based methods, Lecture Notes in Computer Science, 229–245 (2023).

Question Evaluation

Automated/Rule-based Evaluation

- Apply an automated rule-based evaluation system from **Moore et. al**
- Evaluates questions on Item Writing Flaws (IWF) – common mistakes in MCQ generation (ex. unclear questions/answers, answer cues, etc.), compared with human evaluation
- Investigate the potential for an automated generation and evaluation pipeline that combines GPT-4's generated MCQs with **Moore et. al**'s model

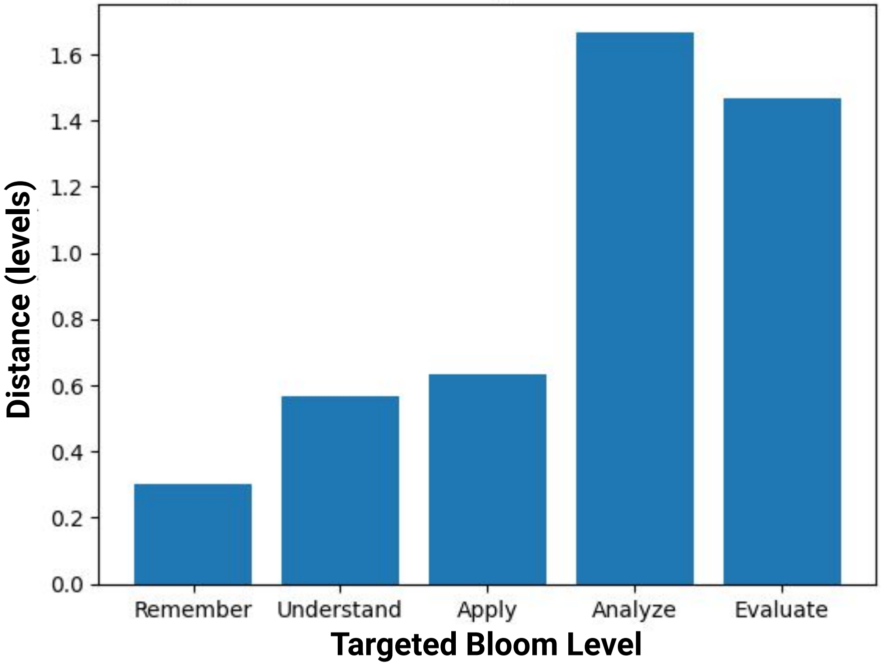
Moore, S., Nguyen, H. A., Chen, T.., Stamper, J., Assessing the quality of multiple-choice questions using GPT-4 and rule-based methods, Lecture Notes in Computer Science, 229–245 (2023).

RQ1: How well does GPT-4 generate MCQs aligned to Bloom's Taxonomy?

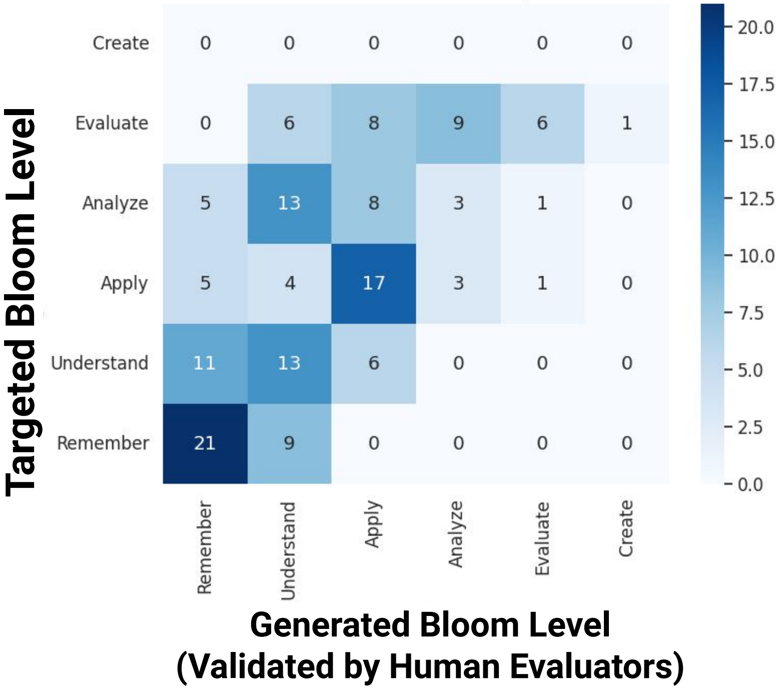
Bloom Taxonomy Levels Alignment: Targeted vs. Generated

Results are aggregated from experiments across all shots configuration

Average Distance Between Targeted Bloom Level & Generated



Confusion Matrix of Generated Questions



	Remember	Understand	Apply	Analyze	Evaluate	Create
Create	0	0	0	0	0	0
Evaluate	0	6	8	9	6	1
Analyze	5	13	8	3	1	0
Apply	5	4	17	3	1	0
Understand	11	13	6	0	0	0
Remember	21	9	0	0	0	0

Spearman correlation coefficient of 0.90 (p=0.037)

82.67%

Questions marked as usable or usable with minor edits.

(by human evaluator, aggregated across all # of shots)

91.33%

Questions marked as relevant or relevant with minor edits.

(by human evaluator, aggregated across all # of shots)

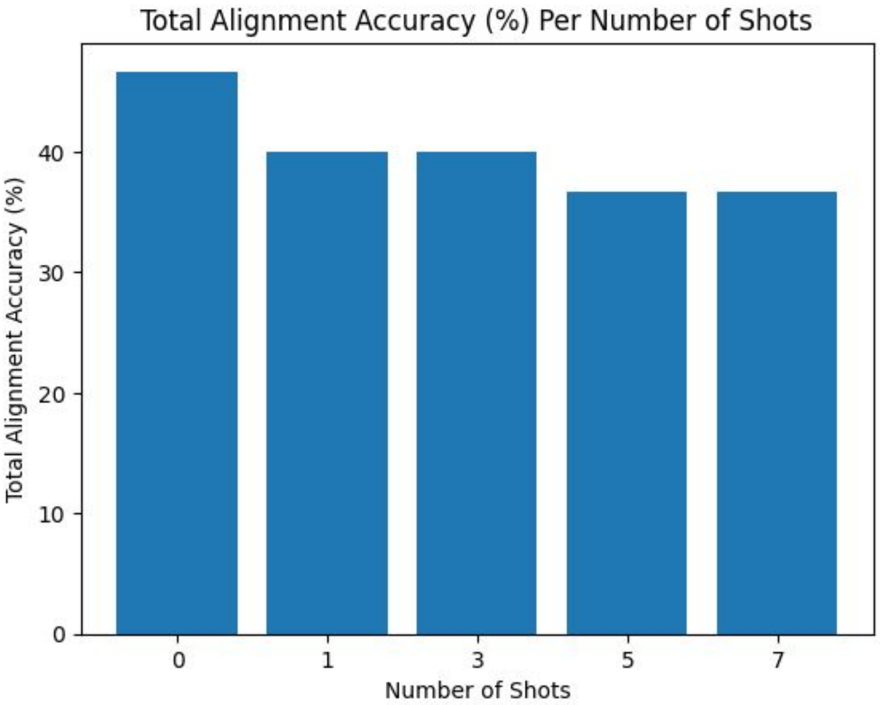
RQ2: How does GPT-4’s performance change with the number of shots/examples given?

How Alignment Varies with # of Shots/Examples?

Alignment accuracy = the % of questions matching the target Bloom level

- As we increase the amount of shots given to GPT-4, performance gets worse

- GPT-4 get confused with more examples -:)



Spearman correlation coefficient of -0.95 (p=0.014)

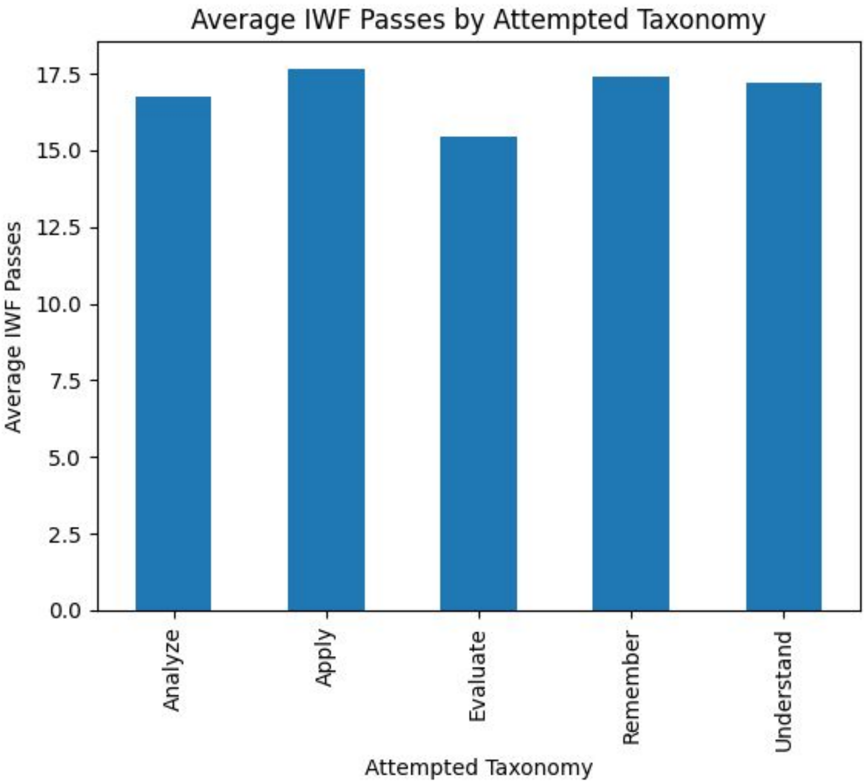
How Relevance & Usability Correlates with # of Shots?

- For all # of shots provided, GPT-4 can reliably generate questions that are usable and relevant

- No obvious relationship is noted with number of shots

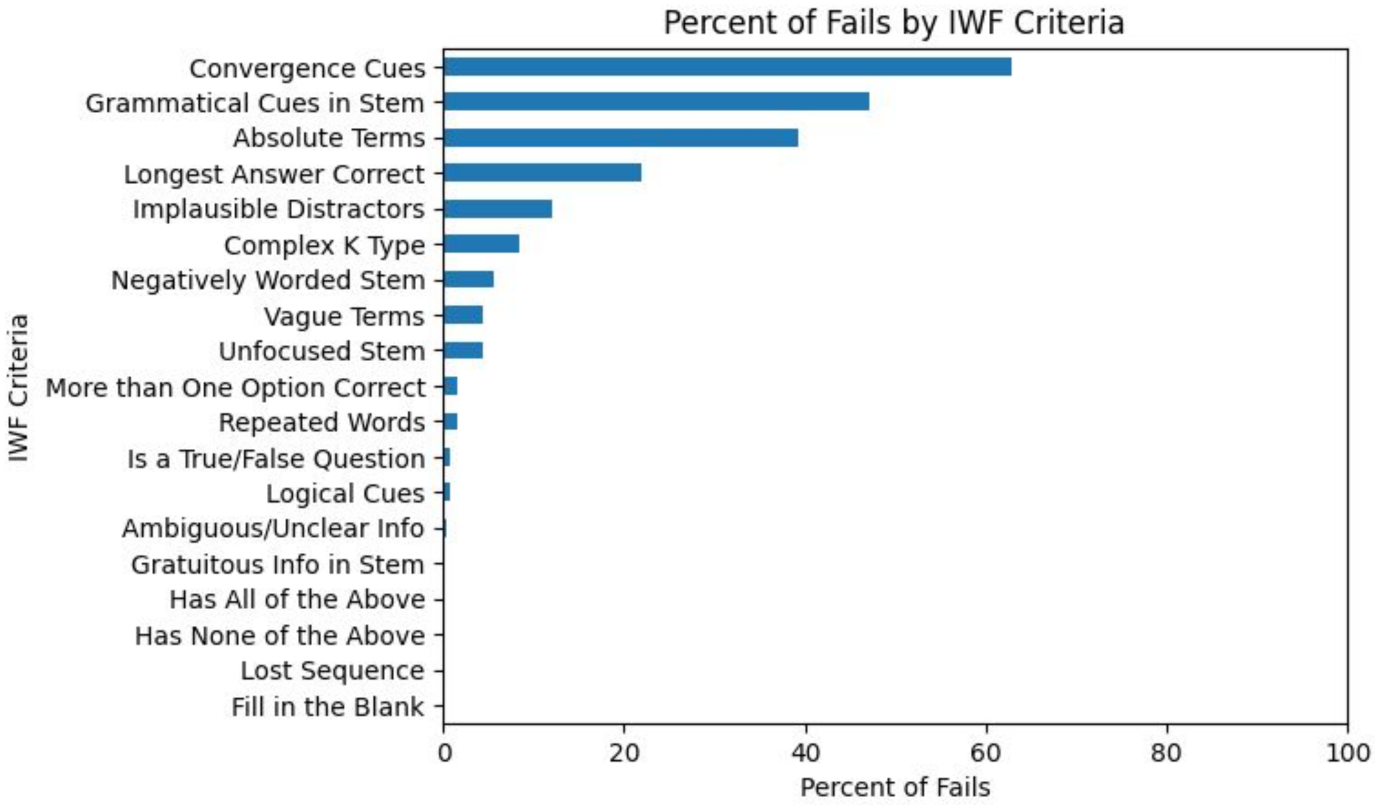
RQ3: How well does automated MCQ evaluation aligned with human feedback on usability ?

Average # of IWF Passes by Attempted Taxonomy



Max # of IWF passes = 19

Percentage of Fails by IWF Criteria



Example of IWF-Convergence Cues

Question: Who is the founding president of UMBC?

Options:

A) Dr. Freeman Hrabowski
B) Dr. Thomas Jefferson
C) Dr. Albin O. Kuhn
D) Dr. Robert Frost

Example of IWF- Grammatical Cues in Stem

Question: A person **who** attends UMBC **is** most likely to be a:

Options:

A) Student
B) Professor
C) Alumni
D) Visitor

IWF assessment vs Human Evaluated Usability

- Dichotomized usability ratings
  - 1-2 → not usable
  - 3-4 → usable
- IWF binarized outcomes
  - >1 flaw → proxy to “not usable”
  - <=1 flaw → proxy to “usable”
- Very little correlation noted
- Our evaluator focuses only on content of MCQs, rather than formats?

43.33% Accuracy
87.79% Precision
37.10% Recall
0.52 F-score

Moore, S., Nguyen, H. A., Chen, T.., Stamper, J., Assessing the quality of multiple-choice questions using GPT-4 and rule-based methods, Lecture Notes in Computer Science, 229–245 (2023).

Take Home Message #1

- GPT-4 can generate questions that are usable & relevant
- Alignment with targeted Bloom level varies, struggle to create questions at higher level
- More work needs to be done improving Bloom's taxonomy alignment, especially at higher levels

Take Home Message #2

- In terms of creating usable & relevant questions, the number of shots do not seem to matter

- More work needs to be done to design better shots/examples so that LLMs understand the Bloom's Taxonomy (e.g. providing contrasting examples from various level)

Take Home Message #3

- IWF rating for MCQ quality does not seem to be aligned well with human experts’ evaluations usability
- Future work needed to better understand the relationship between MCQ quality as measured by IWFs and usability

Practical Implication/Discussion: Instructors? Students?

How can we help instructors?

- Questions at higher levels of Bloom’s Taxonomy encourage greater conceptual understanding **(Jensen et al., 2014)**

- Many question bank don’t include questions on varied levels of Bloom’s Taxonomy **(Swart, 2010; Upahi et al., 2016)**
  - Why? Generating well-designed questions on each level is very time-consuming

How can we help instructors?

- We propose a scalable system for the generation of exam questions across different levels of Bloom's taxonomy

- Manual intervention still needed to ensure alignment to Bloom's taxonomy, however a significant saving compared with generating questions from scratch
    - Manual classification of Bloom level, validation of usability, and relevance for 150 questions took ~90 minutes, i.e. 36 seconds/question

How can we help instructors?

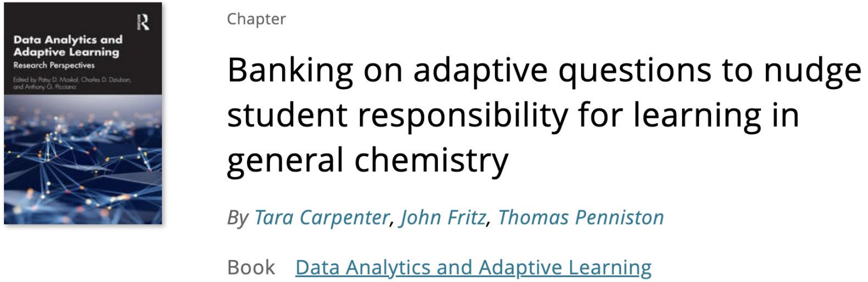
Possible idea to generate more high quality MCQ at higher Bloom level

- Massively oversample for higher-order (Analyze, Evaluate) questions then reject poor questions
  - Insight: Rejection of poor questions is very quick, creating/rephrasing questions is not.

How can we support self-regulated learners?

With additional functionality and enhanced with Learning Analytics

- Used as a **diagnostic tool**
  - Coupled with knowledge tracing, i.e. assessment/tracking of knowledge
  - Help learners understand where they know and what they need to improve
- Used as a **study tool**
  - Coupled with forgetting curve tracking
  - Personalized Spaced-practice tool



Data Analytics and Adaptive Learning
Research Perspectives
Edited by Patsy D. Moskal, Charles D. Dziuban, and Anthony G. Picciano

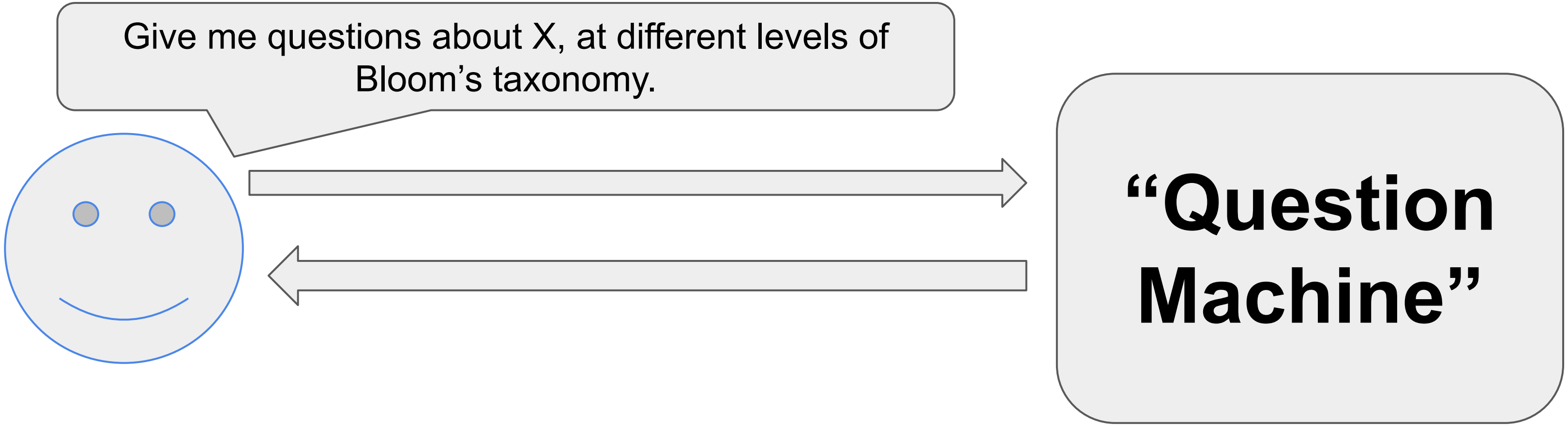
Chapter

Banking on adaptive questions to nudge student responsibility for learning in general chemistry

By *Tara Carpenter*, *John Fritz*, *Thomas Penniston*

Book [Data Analytics and Adaptive Learning]

From a student’s perspective…

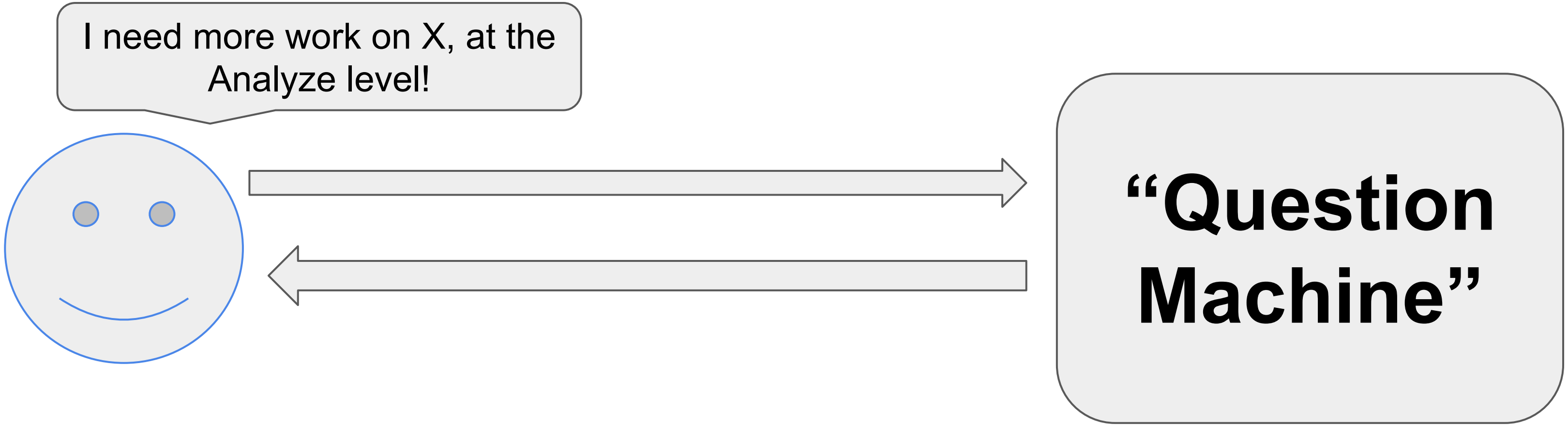


Give me questions about X, at different levels of Bloom’s taxonomy.

“Question Machine”

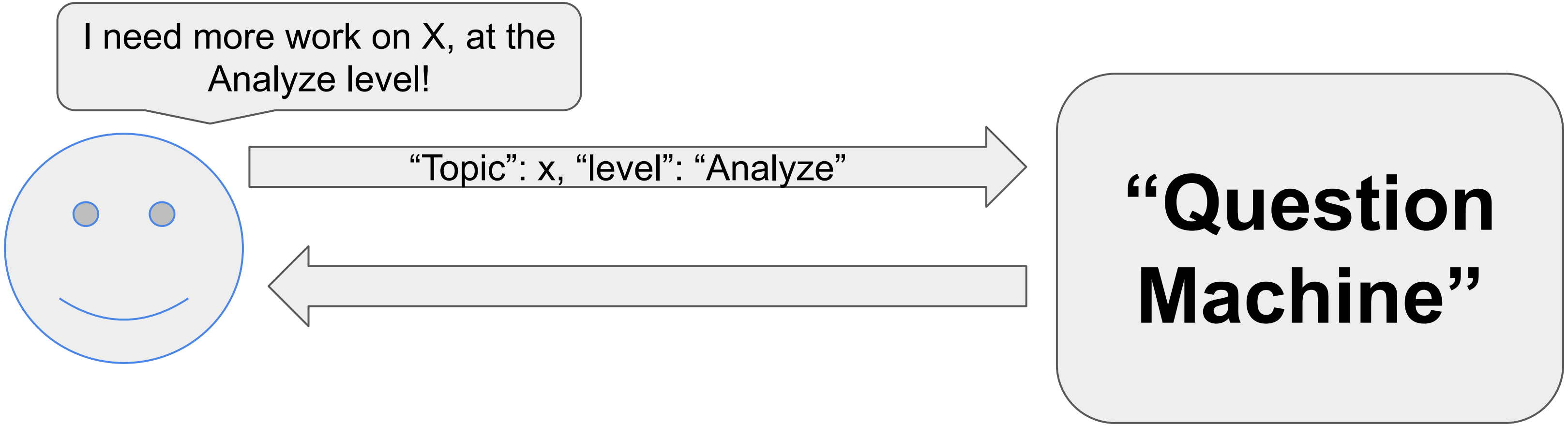
Student uses the diverse initial questions as a diagnostic test to understand what they do not currently know.

From a student’s perspective…



Student reasons about what they do not know and how they can improve based on their performance on the diagnostic test.

From a student’s perspective…



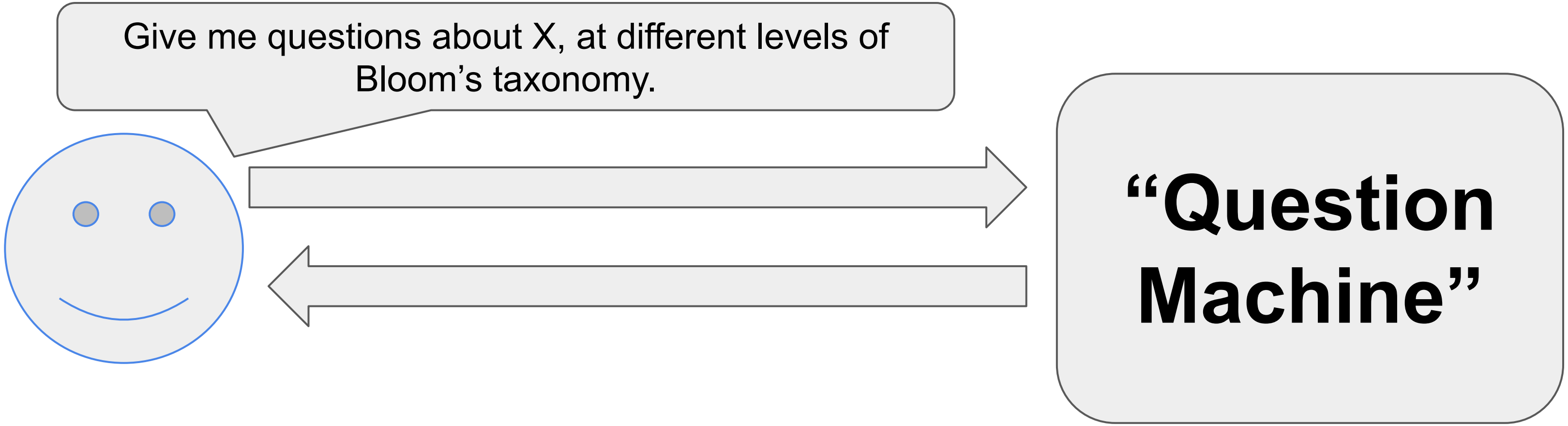
I need more work on X, at the Analyze level!

“Topic”: x, “level”: “Analyze”

“Question Machine”

Student uses more questions to improve on their current knowledge

From a student's perspective…



Give me questions about X, at different levels of Bloom's taxonomy.

"Question Machine"

Student retakes the diagnostic test, creating a loop.

Caveats

- In order to reap benefits, students still need intrinsic motivation to
  - Use the practice tool
  - Productively use, not “gaming”

- Beyond technology: Reward system design? Human psychology?
- Kevin and Kenneth’s insight on this?
  - What is your strategy to keep yourself motivated?
  - “Nudge” your peers to be more self-regulated?

Recap

Key Findings:

- GPT-4 can generate useful and relevant questions, but struggles with aligning with Bloom's taxonomy, especially at higher levels
- Increasing # of shots has no effect on relevance/usability, but has an adverse effect on Bloom's taxonomy alignment
- IWF-based evaluation does not align with human experts' usability evaluation
- Potential usage by instructors and students

Contact Us:

{khwang8265, klwang1008}@inst.hcpss.org

{maryama4, lujiec, choa}@umbc.edu

Thanks & Feedback

http://tiny.cc/133717