Towards AI-Assisted Multiple Choice Question Generation and Quality Evaluation at Scale: Aligning with Bloom's Taxonomy

Kevin Hwang[1], Sai Challagundla[1], Maryam Alomair[2], Lujie Karen Chen[2], Fow-Sen Choa[2]

[1]Glenelg High School, Maryland USA; [2]University of Maryland Baltimore County, Maryland USA

Introduction

1) Multiple Choice Questions (MCQs) are used in education for fast grading and feedback.
2) No current way to automatically generate MCQs based on Bloom's Taxonomy [1] (a taxonomy for the cognitive levels required to answer questions)
3) We automatically generate MCQs based on Bloom's Taxonomy using GPT-3.5.
4) We evaluate generated MCQs using an automated rules-based approach and a domain expert.
5) We assess the alignment of the taxonomy specified in the prompt (GPT-taxonomy) with the CNN-classified taxonomy (ML-taxonomy) and human classified taxonomy (human taxonomy), and the quality of questions based on their prompt-specified taxonomy.

Fig 1 Subplot A (agrmt. between prompt taxonomy and CNN-classified taxonomy) suggests that GPT-3.5 has the capability for generating MCQs based on Bloom's taxonomy: strongest for Evaluation and weakest for Synthesis and Comprehension.

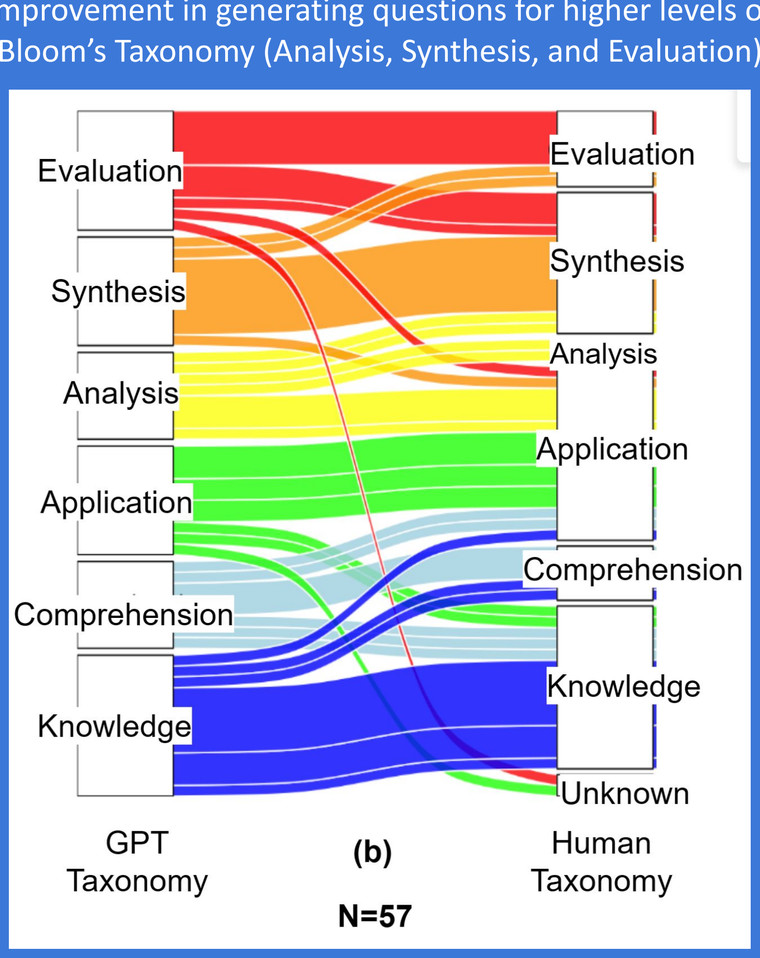


Fig 1 Subplot B (agrmt. between prompt taxonomy and human-classified taxonomy) suggests that there is room for improvement in generating questions for higher levels of Bloom's Taxonomy (Analysis, Synthesis, and Evaluation)

Full Paper and Prompt Template

Methods

Generation:

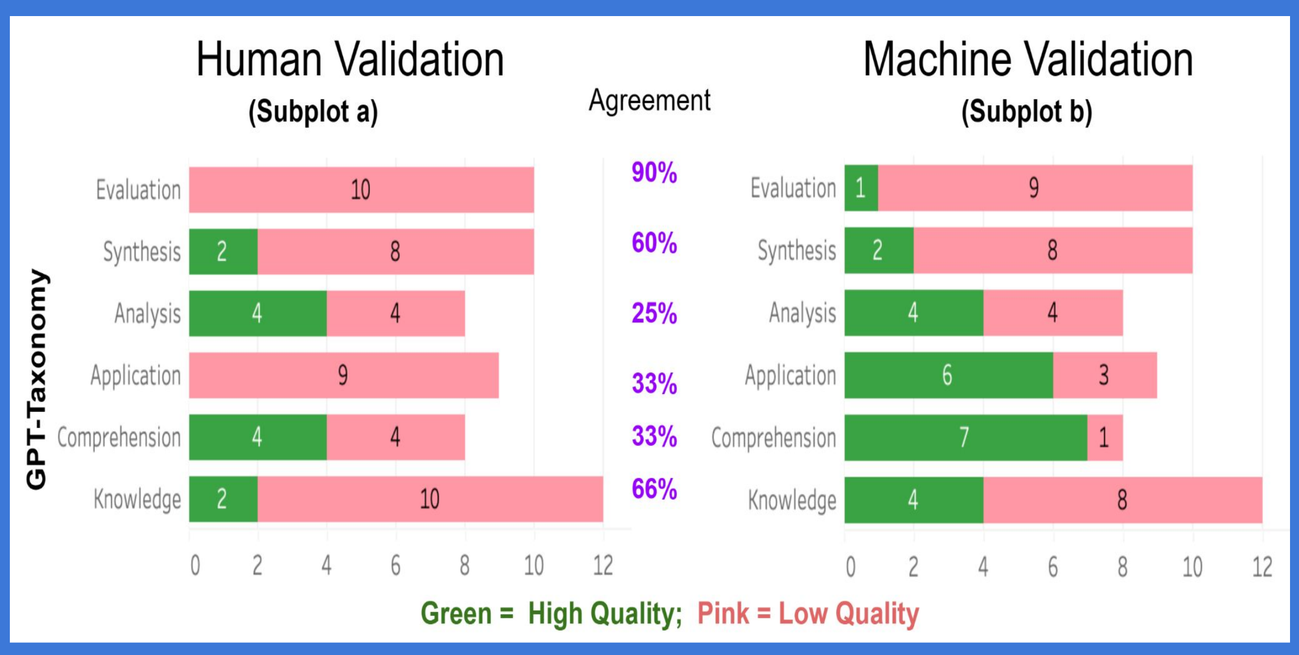
1) Zero-shot prompted (provided no training data) GPT-3.5 to generate MCQs based on textbook excerpts from OpenStax Chemistry 2e and Biology 2e [4]. Prompt template available in QR code.
2) Generated 5 MCQs per every Bloom level, for every excerpt. 120 questions were generated using excerpts from two randomly selected chapters from each textbook.

Evaluation:

1) Used a convolutional neural network (CNN) to verify the Bloom's taxonomy of generated MCQs. Model architecture based on [2].
2) Used the model in [3] to assess the quality of generated questions using IWFs (mistakes in educational MCQs). > 1 IWF = Low Quality, as per [5].
3) Domain expert with 28+ years of STEM educational experience was asked whether they would use a question in a classroom setting and the Bloom's taxonomy the question for each question in a subset of 57 of the MCQs.

Figure 2 shows the # of questions marked as high/low quality for each Bloom level and the percent agreement for each Bloom level.

1) In general, cognitive complexity and generated question quality show an inverse relationship.
2) In general, agreement between humans and machines decreases as cognitive complexity goes down.
3) Domain expert marked 21% of questions as high quality and model marked 42% of questions as high quality, showing difference of perspective.



Take Home Messages

1) GPT-3.5 has promising capabilities in generating MCQs aligning with Bloom's Taxonomy, but aligning better requires further exploration.
2) Particularly at higher levels of Bloom's taxonomy, GPT-3.5 struggles with generating questions that are usable in the classroom, more work needs to be done in refining the quality of questions.
3) Automated question evaluation using IWF and human evaluation differs significantly, requiring exploration on how to best model human validation using automated measures.

References

1) Bloom, B. "A taxonomy of cognitive objectives." *New York: McKay* (1956).
2) Gani, Mohammed Osman, et al. "Bloom's Taxonomy-based exam question classification: The outcome of CNN and optimal pre-trained word embedding technique." *Education and Information Technologies* (2023): 1-22.
3) Moore, Steven, et al. "Assessing the Quality of Multiple-Choice Questions Using GPT-4 and Rule-Based Methods." *European Conference on Technology Enhanced Learning*. Cham: Springer Nature Switzerland, 2023.
4) *OpenStax | Free Textbooks Online with No Catch*. https://openstax.org/. Accessed 24 Sept. 2023.
5) Tarrant, Marie, et al. "The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments." *Nurse Education Today* 26.8 (2006): 662-671.