(a) N=120 — GPT Taxonomy to ML Taxonomy


(b) N=57 — GPT Taxonomy to Human Taxonomy


Generated Multiple Choice Questions

Bloom's Taxonomy Verification: Gani et al Bloom's Taxonomy Classification Convolutional Neural Network

Question Quality Evaluation: Moore et. al Item Writing Flaw Detection Model → > 1 Item Writing Flaw? → True: MCQ is low quality / False: MCQ is high quality


openstax™ → GPT

Bloom's Taxonomy pyramid:
- creating
- evaluating
- analyzing
- applying
- understanding
- remembering

## Human Validation (Subplot a)

| GPT-Taxonomy | | Agreement |
|---|---|---|
| Evaluation | 10 | 90% |
| Synthesis | 2, 8 | 60% |
| Analysis | 4, 4 | 25% |
| Application | 9 | 33% |
| Comprehension | 4, 4 | 33% |
| Knowledge | 2, 10 | 66% |

## Machine Validation (Subplot b)

| GPT-Taxonomy | |
|---|---|
| Evaluation | 1, 9 |
| Synthesis | 2, 8 |
| Analysis | 4, 4 |
| Application | 6, 3 |
| Comprehension | 7, 1 |
| Knowledge | 4, 8 |

Green = High Quality; Pink = Low Quality

Human Evaluator (28+ years educational STEM experience):
1) Would you use this question in a classroom setting?
2) What is the Bloom's taxonomy level of this question?

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