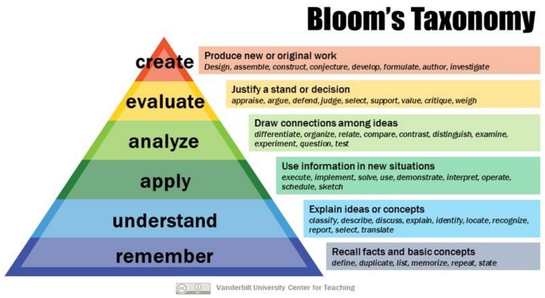
Quality Assessment of LLM-Generated Multiple-Choice Questions in Medical Education: A Preliminary Structured Literature Review

Yash Patel OMS-I[1] Troy Camarata Ph.D.[1]
[1]Baptist Health Sciences University, College of Osteopathic Medicine. Memphis, TN

Introduction

- Multiple-choice questions (MCQs) are widely used in medical education for preclinical learning and board preparation
- Practice testing with MCQs supports learning and improves long-term retention
- High-quality MCQs promote higher-order analytical thinking and more accurate knowledge assessment



Relevance

- Use of large language models (LLMs), such as ChatGPT, is increasing in medical education
- Faculty use LLMs to generate multiple-choice questions (MCQs) for efficiency
- Concerns persist about quality, accuracy, and use without expert oversight
- Because MCQ quality directly impacts learning outcomes, evaluation of LLM-generated content is essential

Objectives

- Examine existing literature on the quality of LLM-generated MCQs in medical education
- Identify common limitations in MCQs, including accuracy, distractor quality, and item-writing flaws
- Identify gaps in student-based evaluation of LLM-generated MCQs
- Inform development of future studies comparing student and faculty evaluation of MCQ quality

Methods

Study Design

- Preliminary structured PubMed literature review
- Keywords: LLMs, MCQ generation, medical education

Inclusion Criteria

- Studies evaluating LLM-generated MCQ Quality
- Medical Education Setting

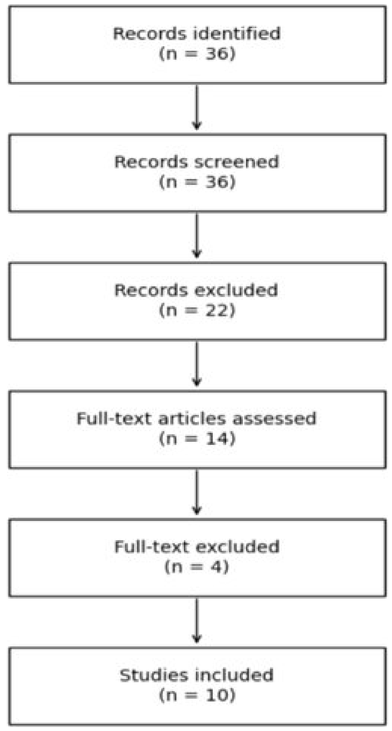
Exclusion Criteria

- Studies on LLMs answering questions
- Editorials or other non-evaluative studies

Data Extraction

- LLM model used
- Error types identified
- Psychometric measures reported
- Study population and level of training

PRISMA Flow Diagram



Results

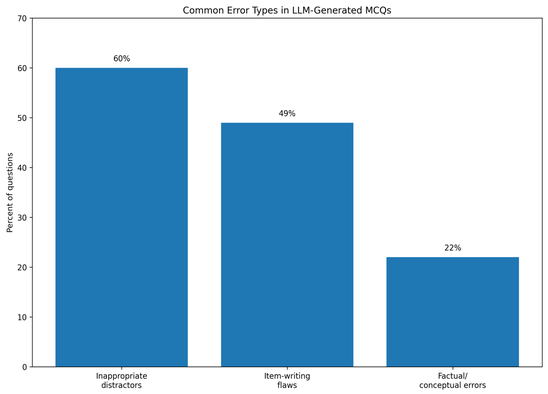
Study Characteristics

- Included studies across preclinical and residency-level medical education
- Multiple LLMs evaluated (e.g., GPT-4, GPT-4o, ChatGPT-o1, DeepSeek R1, Baichuan4)

Key Findings

- Error rates ranged from 22-30%
- Common errors included:
  - Incorrect answer keys
  - Poor Distractor Quality
  - Lack of literature on student evaluation of LLM-generated MCQ quality
  - Consistent recommendation for expert/faculty oversight

Representative Study Findings



Common Error Types in LLM-Generated MCQs. (Camarata et al., 2025)

Conclusion

- LLM-generated MCQs have increased utility in medical education due to their scalability as learning tools to generate topic-specific question items
- Error rates range from 22–30%, with common issues including incorrect answer keys, poor distractor quality, and item-writing flaws
- Expert/faculty oversight is necessary to ensure accuracy and quality
- Limited evidence exists on preclinical student evaluation of LLM-generated MCQs, representing a key gap in the literature and medical education

Future Directions

- Compare medical student vs. faculty evaluation of identical LLM-generated MCQs using a double-blind study design
- Assess evaluation criteria using NBME® and NBOME® question writing rubrics
- Identify differences in recognition of MCQ flaws between students and faculty
- Use findings to develop targeted training frameworks to improve student use of LLMs for MCQ generation

Acknowledgements

I want to thank Dr. Camarata for his mentorship and guidance, and the Baptist Health Sciences University, College of Osteopathic Medicine Office of Research for their support

References

Scan for References