AI IN ASSESSMENT: ETHICS, INNOVATION AND RESEARCH

“

Through transparency, we build trust and ensure accountability.

Victoria Yaneva, PhD,
Manager, NLP Research

NBME®

INTRODUCTION

AI's potential to transform medical education assessment

NBME is exploring the use of AI — more specifically Natural Language Processing (NLP) and machine learning — to provide tailored applications that meet the needs of medical education assessment. NLP explores the processing of natural language (e.g., English, French, Chinese) by computer systems, while machine learning focuses on using algorithms to enable AI to learn from examples. The integration of these innovative technologies with medical education assessment has significant transformative potential, as well as ethical guidelines that must be considered. Two NBME researchers recently discussed the use of AI in assessment.

Presentation Topics

How AI impacts medical education and assessment	3
NBME's ongoing innovative AI research	4
Ethical considerations of AI's use in medical education	6
Building trust	7

Panelists

Dr. Andrea Anderson, MD (Moderator)
Associate Chief and Associate Professor of the Div. of Family Medicine, The George Washington University School of Medicine and Health Sciences

Victoria Yaneva, PhD (Presenter)
Manager, NLP Research

Kimberly Swygert, PhD (Presenter)
Director, Test Development Innovations

HOW DO RECENT ADVANCES IN AI IMPACT MEDICAL EDUCATION?

Many of the recent advances in AI can be traced to a specific type of model called "pretrained transformer models," which have catalyzed breakthroughs across various domains, including ChatGPT or the ability of systems such as Alexa and Siri to accurately recognize speech. Within the realm of medical education, these pretrained transformers offer significant potential enhancements in the following areas:

Automated scoring

Traditionally, adapting AI models to specific exams required significant data training. However, pretrained transformers introduce a notable shift by enabling adaptation even with relatively small datasets through a process called "fine-tuning." This approach leverages the models' existing knowledge from extensive **external** data while incorporating insights from limited **internal** data pertinent to the exam at hand. Evidence shows that this strategy can yield highly accurate automated scoring results.

Test development

AI models can help obtain preliminary estimates of item difficulty and expected response times prior to conducting pretests, a capability that enhances fairness in testing. AI-predicted response times can help in equitable assignment of newly developed items to test forms, so that examinees have comparable amounts of time left to respond to the live terms.

Item development

Using generative AI to assist human experts in creating test items is an area that has generated particular excitement. This approach is already applied in language testing, with some exams relying on AI-generated reading passages. However, one important difference between reading passages and clinical test items is that the latter cannot be fictional. Instead, items in medical education must maintain clinical correctness and relevance to current medical practices.

This requirement introduces a unique challenge, as the use of generative AI in content creation raises concerns about the potential generation of factually inaccurate information, leading to "hallucinations" in the generated content. Addressing this issue is imperative to ensuring the reliability and relevance of AI-assisted content development in medical education. Numerous approaches to limiting hallucinations are being investigated by the AI community.

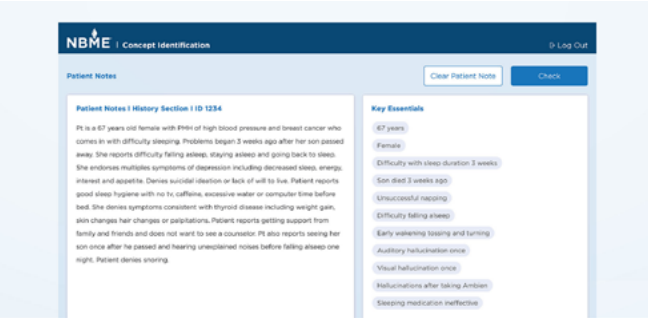
ONGOING AI RESEARCH AT NBME

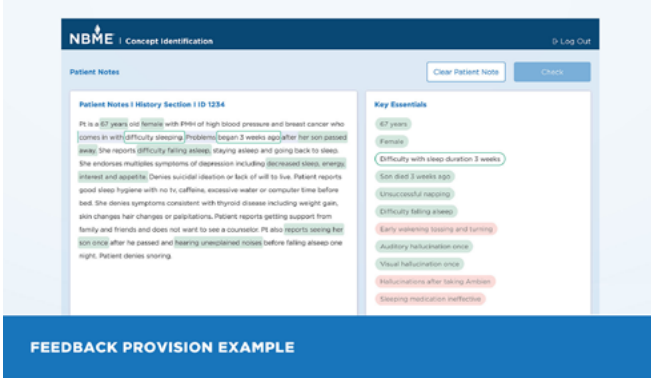
NBME is investigating the use of AI to develop innovative capabilities specifically tailored to the needs of medical education assessment. This innovation is developed through research, collecting evidence and presenting for peer review. Our current work spans the following areas:

Enabling the measurement of complex constructs

To evaluate complex constructs like communication skills and clinical reasoning, medical educators require innovative assessments that go beyond the multiple-choice format. However, such assessments require the ability to efficiently score a vast number of free-text responses. For example, if 1,000 examinees see 20 open-ended items of any kind that require free-text responses, this would result in 20,000 responses to score.

Free-text responses can vary significantly in terms of phrasing, and we have evidence that AI approaches can manage this variability at scale, reserving human review for a smaller fraction of the more challenging responses.

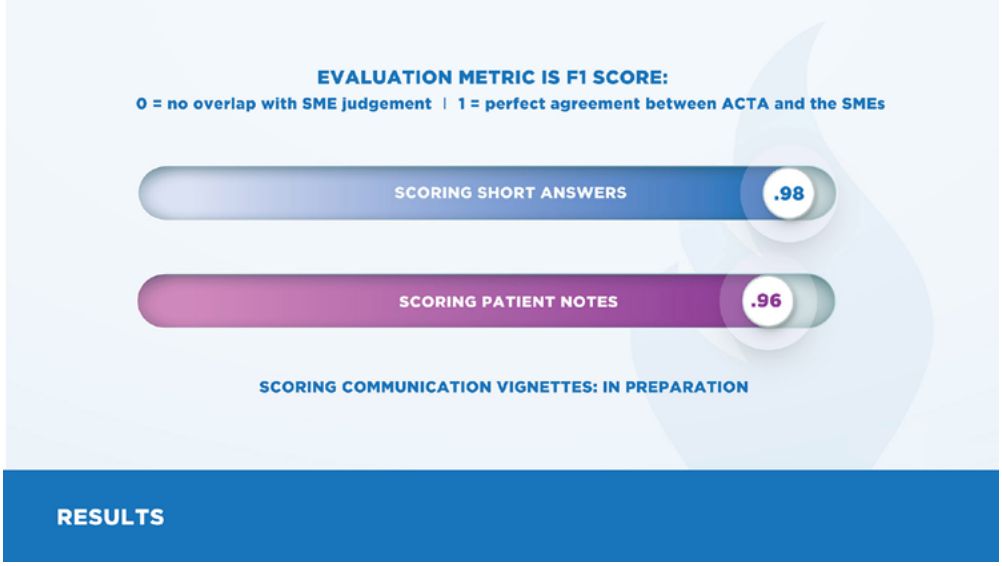




Shown in the graphics above, the highlights indicate specific text within an examinee-written patient note that corresponds to a key concept within the scoring rubric, as identified by the automated scoring system. Key concepts from the rubric highlighted in red were ones that were not found within the patient note.

ONGOING AI RESEARCH AT NBME (CONT.)

The process begins with subject matter experts (SMEs) who review a large sample of free-text responses and mark them as correct or incorrect. This information is then used to fine-tune transformer models. The predictions (see Scoring Example below) of the model are compared to those of the SMEs, and the resulting metric is called an F score. The F score is a measure of the extent to which the scores assigned by the model overlap with those assigned by human raters, as illustrated below.



Providing personalized feedback that facilitates learning

Another example is the need to provide personalized feedback to help learner growth in areas such as developing communication skills. If this feedback is not accurate or prompt, learners may not benefit from it. Approaches similar to those used in automated scoring can be applied to this problem and, by learning from many examples created by SMEs, can help deliver accurate and personalized feedback almost immediately.

Generating new items

NBME has begun an exploration of AI-assisted item development; however, it is still in its early stages. A key requirement for the success of this initiative is to ensure the items developed through this innovative, efficient process meet the same rigorous quality standards as current items. Achieving this necessitates placing human experts at the center of this process, enabling them to steer development and uphold our quality benchmarks.

ETHICAL CONSIDERATIONS

Though the field of ethical AI has grown rapidly along with its technological advances, at the center of it remains machines and algorithms that use data. The responsibility for the ethical use of these applications rests entirely on humans. Ethical guidelines have evolved along with the technology and typically fall into six main categories.

Data Privacy

Incorporate clear rules and assurances of how data will be collected, how personal data will be protected, how long data will be retained, and whether it's acceptable to share or repurpose data.

Accountability

Consider regulatory requirements regarding the development and deployment of AI technologies to ensure compliance. Be prepared to explain the aims, motivations and reasons for using AI.

Fairness

Begin with the assumption that the use of AI is never neutral or impartial, and acknowledge that the use of AI, especially for tasks like automated decision-making, could lead to discriminatory outcomes.

Transparency

Prepare to share information and an explanation about how AI is used through methods appropriate to your audience.

Human Control and Oversight

Maintain human oversight to ensure that an organization's use of AI systems is transparent, explainable and trustworthy.

Promotion of Human Values

Support the principle of non-maleficence, or "do not harm," applying to both foreseeable and negative outcomes"

BUILDING TRUST

It is important to proactively build trust regarding the use of AI. At NBME, this includes:

- Creating multidisciplinary teams to collect robust evidence and present it to the public through peer-reviewed research. Findings are communicated to both the medical education and AI research communities to ensure that the right questions are asked by the right experts.
- Releasing datasets for the purpose of competitions or collaborations, out of which come solutions that are available to the community for scrutiny. Through transparency, we build trust and ensure accountability.

Our commitment to validity and ethics

NBME is working to identify item-writing tasks and process that must continue to rely on human expertise, and those can be made more efficient through the use of AI. NBME is committed to transparency with the medical education community, and as our work to develop a set of ethical guidelines moves forward, we will continue to provide updates and information.

AI is a tool that will be integrated into NBME's existing foundational measurement work; it will not replace it. We will design our future ethical guidelines to merge with existing guidelines for test development, test scoring, and, in particular, the gathering of evidence for test validity. We will address potential AI bias with the same careful, methodical approach we use to tackle other potential threats to validity.

FURTHER READING

Click through to learn more.

- NBME's Kaggle competition on Automated Scoring of Clinical Patient Notes
- NBME's Research Library
- NBME NLP Conference Proceedings
- The BEA'24 Shared Task on Automated Prediction of Item Difficulty and Item Response Time
- European Commission's Ethics Guidelines for Trustworthy AI
- President Biden Executive Order on the Safe, Secure and Trustworthy Development and Use of Artificial Intelligence
- *The Ethics of AI Ethics – An Evaluation of Guidelines*