COMMENTARY

Leveraging Large Language Models to Advance Certification, Physician Learning, and Diagnostic Excellence

Ting Wang, PhD, David W. Price, MD, and Andrew W. Bazemore, MD, MPH

Diagnostic errors are a significant challenge in health care, often resulting from gaps in physicians' knowledge and misalignment between confidence and diagnostic accuracy. Traditional educational methods have not sufficiently addressed these issues. This commentary explores how large language models (LLMs), a subset of artificial intelligence, can enhance diagnostic education by improving learning transfer and physicians' diagnostic accuracy. The American Board of Family Medicine (ABFM) is integrating LLMs into its Continuous Knowledge Self-Assessment (CKSA) platform to generate high-quality cloned diagnostic questions, implement effective spaced repetition strategies, and provide personalized feedback. By leveraging LLMs for efficient question generation and individualized learning, the initiative aims to transform continuous certification and lifelong learning, ultimately enhancing diagnostic accuracy and patient care. (J Am Board Fam Med 2025;38:599–602.)

Keywords: **Artificial Intelligence, Certification, Continuing Education, Diagnostic Errors, Examination Questions, Family Medicine, Formative Feedback, Large Language Models, Physicians, Self-Assessment**

Diagnostic errors represent a critical and costly challenge in health care, contributing to significant patient harm and financial burdens. These errors account for approximately 10% of patient deaths and are the leading cause of medical malpractice claims,[1,2] with nearly 795,000 Americans experiencing severe harm annually due to diagnostic errors.[3] Factors such as gaps in physicians' knowledge[4] and miscalibration between confidence and diagnostic accuracy[5,6] contribute significantly to these errors. This article explores how large language models (LLMs), a subset of artificial intelligence (AI), could bridge critical gaps in diagnostic education by improving learning transfer and enhancing physicians' diagnostic accuracy. We also review a certifying board's efforts to test these possibilities.

The Problems and Current Limitations

Diagnostic errors stem from many complex causes, including gaps in physician knowledge, cognitive biases, and misalignment between diagnostic confidence and accuracy.[4–6] Medical school and residency training, continuing medical education, and ongoing certification have not sufficiently addressed these gaps. Member boards of the American Board of Medical Specialties (ABMS), including the American Board of Family Medicine (ABFM), have sought to tackle these challenges by implementing longitudinal assessment platforms. These platforms provide continuous assessments to help physicians enhance their medical knowledge and diagnostic skills.

Research has shown that spaced repetition—revisiting material at increasing intervals—is highly effective for retaining knowledge.[7–9] However, current spaced repetition strategies often focus on repeating the same questions, which may not effectively capture the complexity and variability of real-world diagnostic tasks. Cloned multiple-choice questions, designed to test the same knowledge domain but in different clinical contexts, may offer a more robust opportunity for

This article was externally peer reviewed.
Submitted 23 October 2024; revised 2 January 2025; accepted 21 January 2025.
From the American Board of Family Medicine, Lexington, KY (TW, DWP, AWB); Department of Family Medicine, University of Colorado School Anschutz School of Medicine, Aurora, CO (DWP).
Funding: The authors accepted funding from ABMS and the Gordon and Betty Moore Foundation.
Conflict of interest: None.
Corresponding author: Ting Wang, PhD, 1648 McGrathiana Parkway, Suite 550, Lexington, KY 40511 (E-mail: twang@theabfm.org).

J Am Board Fam Med: first published as 10.3122/jabfm.2024.240385R1 on 25 August 2025. Downloaded from http://www.jabfm.org/ on 6 November 2026 by guest. Protected by copyright.

learning transfer. They encourage physicians to apply their knowledge in varied scenarios, potentially better supporting diagnostic improvement compared with mere repetition of the same questions.[9] However, the labor-intensive nature of human item generation has long limited the scalability of these strategies—a challenge that AI, particularly LLMs, are well-positioned to address.

LLM-Driven Initiatives

With support from the ABMS and the Gordon and Betty Moore Foundation, the ABFM is actively exploring how LLMs can be integrated into longitudinal assessment platforms to support diagnostic education. This initiative focuses on 3 areas: improving the efficiency of question generation, enhancing learning transfer through spaced repetition of different clones, and providing personalized feedback to physicians. By generating multiple iterations of questions that assess and reinforce the same skill or knowledge, we aim to understand how spaced repetition improves diagnostic accuracy, particularly when knowledge needs to be applied in varied clinical situations.

Project Goals

Our project aims are outlined in the following.

Test the Practicality and Psychometric Soundness of LLM-Generated Clone Items

Developing high-quality diagnostic questions requires significant human input. ABMS member boards, including ABFM, invest considerable resources in building robust item banks for continuous certification assessments. We aim to evaluate whether LLMs can efficiently generate cloned diagnostic questions that maintain the psychometric integrity of the originals, thereby reducing the human resource burden while maintaining quality. We do not anticipate that human input will be eliminated, but that it can be focused on fine tuning question nuance rather than developing items de novo.

Quantify the Impact of Spaced Repetition on Learning Transfer

By comparing spaced repetition outcomes with original and cloned items, we will explore which repetition intervals and frequencies maximize learning transfer. This has direct implications for improving diagnostic accuracy, as repeated exposure to varied diagnostic scenarios may help reduce errors.

Assess the Impact of Personalized Feedback

We will evaluate the technical feasibility, user feedback, and impact of personalized feedback generated by LLMs on physicians' accuracy and confidence. Prior studies have shown that personalized feedback is more effective in promoting engagement and improving learning outcomes than standardized feedback.[10,11] Leveraging LLMs for tailored feedback could enhance metacognitive skills, helping physicians recognize strengths and areas for improvement.

Approach

To achieve these aims, we are using ABFM's Continuous Knowledge Self-Assessment (CKSA) platform and its extensive item bank to train LLMs. The CKSA platform presents participants with multiple-choice questions and provides immediate feedback. Throughout the study, LLM-generated cloned questions will be introduced at various intervals, allowing us to assess the impact of different spaced repetition strategies on learning transfer.

Our study design involves three important steps.

Question Generation and Calibration

LLMs will generate cloned items for each original diagnostic question. The original questions have gone through a rigorous review process to eliminate potential bias. Clones will be evaluated by human reviewers for alignment with the original objectives, potential introduction of new biases, and calibrated for difficulty using psychometric models. We aim to generate sets of cloned items with varying degrees of modification that match the difficulty and objectives of the original questions. We will also document the operational challenges and share successful practices regarding human input in training and verifying LLM outputs.

Spaced Repetition Experimental Design

Participants will be divided into control and experimental groups, with experimental groups receiving varied repetitions of cloned items over multiple quarters. We will measure the impact of these repetitions on both diagnostic accuracy and confidence.

J Am Board Fam Med: first published as 10.3122/jabfm.2024.240385R1 on 25 August 2025. Downloaded from http://www.jabfm.org/ on 6 November 2026 by guest. Protected by copyright.

Personalized Feedback

We will also examine the end user acceptance and impact of personalized feedback generated by LLMs, tailored to individual performance, on engagement, learning outcomes, and confidence improvement, with the goal of fostering active participation and enhancing diagnostic skills. Participants who opt-in to the study will be randomly assigned to either a control or an experimental group. Those in the control group will receive standardized, static feedback in the current CKSA format. Participants in the experimental group will receive a prompt such as, "Would you please share your rationale for choosing X?" (tailored to their specific choice). Following their response, the LLM will provide personalized critiques, addressing the reasoning shared by the participant and offering or reinforcing the correct response's rationale. This iterative, conversational approach is designed to stimulate deeper reflection and promote active learning by adapting to each participant's input, challenging their assumptions, and supporting a more thorough understanding. Importantly, participants in the experimental group will have the flexibility to terminate their interaction with the LLMs at any point during the process. Physicians who decline to participate in the study will still be able to participate in CKSA or other certification activities of their choosing.

The ABFM operates under an existing research protocol approved by the American Academy of Family Physicians Institutional Review Board (IRB), which permits the use of routine organizational data—collected as part of normal ABFM operations—to study the effects of certification activities on physicians' knowledge, skills, and performance. This study builds on previous research, similar in scope and methodology, conducted under this IRB approval. We anticipate that the current study will also fall within the bounds of the existing IRB approval. If the use of LLMs raises concerns, we will promptly seek supplemental IRB approval to ensure the research remains compliant with all necessary guidelines.

Why This Matters

The findings from this project have the potential to transform how ABMS member boards approach continuous certification and lifelong learning. By demonstrating the efficacy of LLMs in generating high-quality assessment items and personalized feedback, we could significantly reduce the costs and time associated with maintaining robust item banks. More importantly, incorporating cloned items in spaced repetition offers a potential pathway to more effective knowledge transfer, which could ultimately lead to improved diagnostic accuracy and better patient outcomes.

Furthermore, our work on personalized feedback addresses a critical gap in current clinical practice—the lack of timely, meaningful feedback on diagnostic performance. By providing individualized feedback through LLM-powered systems, we aim to foster a more reflective learning environment that supports continuous professional development. In alignment with the missions of ABMS and ABFM, this project seeks to enhance the value of continuous certification by creating a more personalized, efficient, and effective learning process.

Conclusion

Integrating LLMs into medical education and assessment represents a pivotal advancement in enhancing physician learning and competency assessment. Through LLM-driven question generation, spaced repetition, and personalized feedback, we aim to advance both the quality and effectiveness of assessment platforms, with the goal of improving diagnostic accuracy and patient care. The scalability and precision offered by LLMs present a promising future for medical education and certification, aligning with the broader goals of ABFM and ABMS to support lifelong learning and professional development. The lessons learned from implementing LLMs will be valuable to organizations involved in lifelong learning and professional development, offering insights for those willing to adapt in the AI era to achieve their missions more effectively and efficiently.

To see this article online, please go to: http://jabfm.org/content/38/3/599.full.

References

1. National Academies of Sciences, Medicine. Health Care. In: *Implementing Strategies to Enhance Public Health Surveillance of Physical Activity in the United States*. National Academies Press (US); 2019. Accessed December 23, 2024. Available at: https://www.ncbi.nlm.nih.gov/books/NBK545645/.
2. Henriksen K, Dymek C, Harrison MI, Brady PJ, Arnold SB. Challenges and opportunities from the Agency for Healthcare Research and Quality (AHRQ)

J Am Board Fam Med: first published as 10.3122/jabfm.2024.240385R1 on 25 August 2025. Downloaded from http://www.jabfm.org/ on 6 November 2026 by guest. Protected by copyright.

research summit on improving diagnosis: a proceedings review. Diagnosis (Berl) 2017;4:57–66.
3. Newman-Toker DE, Nassery N, Schaffer AC, et al. Burden of serious harms from diagnostic error in the USA. *BMJ Quality & Safety*. Published online 2023. Accessed November 25, 2023. Available at: https://qualitysafety.bmj.com/content/early/2023/07/16/bmjqs-2021-014130?versioned=true.
4. Sarkar U, Bonacum D, Strull W, et al. Challenges of making a diagnosis in the outpatient setting: a multi-site survey of primary care physicians. BMJ Qual Saf 2012;21:641–8.
5. Cifu AS. Diagnostic errors and diagnostic calibration. JAMA 2017;318:905–6.
6. Cassam Q. Diagnostic error, overconfidence and self-knowledge. Palgrave Commun 2017;3:1–8.
7. Butler AC. Repeated testing produces superior transfer of learning relative to repeated studying. J Exp Psychol Learn Mem Cogn 2010;36:1118–33.
8. Pan SC, Rickard TC. Transfer of test-enhanced learning: meta-analytic review and synthesis. Psychol Bull 2018;144:710–56.
9. Price DW, Wang T, O'Neill TR, et al. The effect of spaced repetition on learning and knowledge transfer in a large cohort of practicing physicians. Academic Medicine. Published online 2023;10–1097.
10. Tan CH, Lee SS, Yeo SP, Ashokka B, Samarasekera DD. Developing metacognition through effective feedback. Med Teach 2016;38:959–959.
11. Karaoglan Yilmaz FG, Yilmaz R. Learning analytics intervention improves students' engagement in online learning. Tech Know Learn 2022;27:449–60.

J Am Board Fam Med: first published as 10.3122/jabfm.2024.240385R1 on 25 August 2025. Downloaded from http://www.jabfm.org/ on 6 November 2026 by guest. Protected by copyright.