# Intersection of Artificial Intelligence and Medical Education (Student Abstract)

**Keefer P. Wu[1], Patricia C. Tsang[2]**
[1]Northeastern University
[2]MedStar Health
wu.kee@northeastern.edu, patricia.c.tsang@medstar.net

**Abstract**

Can advanced AI-driven technologies transform the traditionally arduous educational process in medicine? This study takes a deep dive into how the publicly available OpenAI ChatGPT-3.5 performs in answering board-style questions designed for physicians training to become pathologists. Correctly answering 75% of 543 questions using an engaging and fast-paced format was an impressive performance. It underscores the potential as well as improvement opportunities of using interactive AI in future medical training.

## Introduction

Medical education is a long and arduous process that typically requires 7 to 10 years of intense coursework and practical training. Pathology is a diagnostic medical specialty that generally encompasses tissue diagnosis of all organ systems (surgical pathology) as well as laboratory testing of blood and body fluids (clinical pathology). This study examines how a publicly available large language model (LLM), OpenAI's ChatGPT-3.5, performs in answering pathology board-style review questions designed for post-doctorate pathologists–in-training. Incorrect answers were analyzed and categorized. Understanding the capabilities and limitations of ChatGPT can shed light on the role that AI-driven technologies may potentially play in shaping the future of medical education.

## Method

Board-style multiple-choice review questions (n=543) were randomly selected from several pathology online question bank sources. These questions were entered one by one into ChatGPT-3.5, and responses were recorded and analyzed. Incorrect responses were categorized into error types, and facts were checked against the medical literature using PubMed searches. The topics spanned 5 major Pathology disciplines. Since ChatGPT did not recognize images, questions containing relevant images were excluded from the study.



## Results

ChatGPT-3.5 correctly answered 406 of the 543 (75%) questions. Answers typically appeared within 1-2 seconds for each question, followed by an explanation. Among the major disciplines of Pathology (Table 1), its performance was the best in informatics & digital pathology (95% correct), followed by molecular pathology (91% correct). Similar scores were observed in surgical pathology, clinical pathology and hematopathology (70% to 73%).

| DISCIPLINE | CORRECT | TOTAL |
|---|---|---|
| **Surgical Pathology** | **204 (73%)** | **278** |
| **Clinical Pathology** | **101 (71%)** | **143** |
| **Hematopathology** | **37 (70%)** | **53** |
| **Informatics & digital pathology** | **35 (95%)** | **37** |
| **Molecular pathology** | **29 (91%)** | **32** |
| **OVERALL** | **406 (75%)** | **543** |

Table 1: Performance Score of ChatGPT on Answering Questions in Various Pathology Disciplines

### Error Categories

Incorrect answers, totaling 137 (25%), could be categorized into 5 error types**:**

- Factual inaccuracy or AI hallucination was the most common error type (41%), such as the deletion of 3 α genes in alpha thalassemia being mistaken as "hemoglobin Bart's hydrops fetalis" (correct: hemoglobin H disease).

- Not recognizing the significance of the laboratory values provided, such as 4.0 being an abnormally low urine pH value even when the normal range was provided (20%).
- Errors associated with questions requiring rank ordering of the most useful, most important or most likely answer choice (17%).
- Errors in patient management decisions, requiring judgment given certain clinical scenarios (16%), e.g. recommending bone marrow biopsy in a patient who is more likely to have reactive increase in white blood cells than a primary marrow disorder.
- Not recognizing the limitations of laboratory testing, e.g. false negative results, inferring incorrectly that a spurious laboratory result was a part of the disease process (6%).

Some of the responses included a disclaimer, such as:

> It's important to consult with a qualified medical professional for accurate diagnosis and management of any hematologic disorder, as these conditions can be complex and require specialized evaluation.

## Discussion

The questions garnered in this study are used by physicians who are training to become specialists in pathology after medical school. They are designed to help pathologists-in-training prepare for their certifying board examinations after 4 years of residency (Jacobs et. al. 2023).

An overall score of 75% by ChatGPT-3.5 on the board-style review questions was impressive, probably equivalent to an early-career physician who has completed several years of pathology specialty training. Since it is unclear how closely this set of questions accurately reflected the scope and difficulty of the official pathology board certifying examinations, no conclusion could be drawn whether this exact performance would meet the passing threshold.

However, the rapid speed with which each question was processed and answered by the AI program (within 1 to 2 seconds) clearly exceeded normal human mental capacity. The explanations that followed contributed to an engaging, interactive and fast-paced learning experience (Lee 2023).

AI hallucinations, the predominant error type observed (53%), are defined as answers generated by AI models that are confidently presented as facts but are objectively incorrect (Sallam 2023). There are multiple reasons to explain how hallucinations may occur, such as low quality of the training data, use of idioms/slang, and limited datasets (Athaluri et. al. 2023). Since AI models are simply a product of the training data available, any input that is biased or skewed can result in output error.

OpenAI gathers data from multiple sources, including Wikipedia and publicly available websites, articles, and online books. Many of these datasets are not checked for misinformation or biases as a tradeoff for a sizably larger amount of training data that can be assimilated (Sallam 2023). This method of data collection is expected to increase the probability of inaccuracies, especially in niche areas with limited datasets.

This study also demonstrated other limitations of using LLM to answer questions, such as decoding diagnostic images, and predicting the best course of action for patient management in settings where medicine is an art rather than a science. The process of interpreting a given laboratory value as normal or abnormal, and inferring its significance has also proven to be a challenge for ChatGPT-3.5. Overall, the AI model was more reliable in answering questions at face value related to disease processes than deciphering the less likely scenario of spurious laboratory results for which public available data tended to be limited.

## Conclusion

In conclusion, AI-driven technology may offer an engaging tool to complement the traditional model of medical education. This study suggests the potential of LLM to revolutionize medical education by making it more interactive and fast-paced. It also underscores the need for continued refinement and improved reliability of AI-generated information. LLM output data, while powerful and fast, are not immune to biases or errors that may be woven into public sources. This vulnerability becomes particularly apparent in niche areas with limited public data. Perhaps future efforts can focus on using curated and current medical information as LLM training sources. This could be accomplished by close collaboration among AI developers, medical experts, and educators, which would be crucial for harnessing the full potential of AI to shape the future of medical education.

## References

Athaluri, S. A., Manthena, S. V., Kesapragada, V. S. R. K. M., Yarlagadda, V., Dave, T., & Duddumpudi, R. T. S. 2023. Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. *Cureus*, 15(4), e37432. doi.org/10.7759/cureus.37432.

Jacobs, J. W., Booth, G. S., Usmani, A., Burner, J., & Adkins, B. D. 2023. Fellowship Board Pass Rates Rising: Analysis of Pathology Subspecialty Board Examination Performance. *Archives of pathology & laboratory medicine* 147(8), 964–968. doi.org/10.5858/arpa.2022-0129-OA.

Lee H. 2023. The rise of ChatGPT: Exploring its potential in medical education. *Anatomical sciences education* 10.1002/ase.2270. doi.org/10.1002/ase.2270.

Sallam M. 2023. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare (Basel)* 11(6), 887. doi.org/10.3390/healthcare11060887.