Performance of ChatGPT in Israeli Hebrew Internal Medicine National Residency Exam

David J. Ozeri MD[1], Adiel Cohen MD[4], Noa Bacharach MD[5], Offir Ukashi MD[3], and Amit Oppenheim MD[2]

[1]Division of Rheumatology, [2]Department of Medicine A, and [3]Gastroenterology Institute, Sheba Medical Center, Tel Hashomer, Israel
[4]Department of Obstetrics and Gynecology, Hadassah Medical Center, Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel
[5]Department of Pediatrics, Leumit Health Services, Tel Aviv, Israel

ABSTRACT **Background:** Completing internal medicine specialty training in Israel involves passing the Israel National Internal Medicine Exam (*Shlav Aleph*), a challenging multiple-choice test. multiple-choice test. Chat generative pre-trained transformer (ChatGPT) 3.5, a language model, is increasingly used for exam preparation.

Objectives: To assess the ability of ChatGPT 3.5 to pass the Israel National Internal Medicine Exam in Hebrew.

Methods: Using the 2023 *Shlav Aleph* exam questions, ChatGPT received prompts in Hebrew. Textual questions were analyzed after the appeal, comparing its answers to the official key.

Results: ChatGPT 3.5 correctly answered 36.6% of the 133 analyzed questions, with consistent performance across topics, except for challenges in nephrology and biostatistics.

Conclusions: While ChatGPT 3.5 has excelled in English medical exams, its performance in the Hebrew *Shlav Aleph* was suboptimal. Factors include limited training data in Hebrew, translation complexities, and unique language structures. Further investigation is essential for its effective adaptation to Hebrew medical exam preparation.

IMAJ 2024; 26: 86–88

KEY WORDS: chat generative pre-trained transformer (ChatGPT), exam, Hebrew, internal medicine, residency

In the state of Israel completing specialty training in internal medicine requires, among other things, passing the Israel National Internal Medicine Exam (*Shlav Aleph*), a challenging multiple-choice test. Preparing for this exam is considered one of the greatest challenges of becoming a specialist in internal medicine. Many residents take time to prepare for the exam. The resources utilized to pass the exam are exam style multiple choice question banks and studying from *Harrison's Principles of Internal Medicine*, and relevant professional expert guidelines.

The exam is composed of 150 multiple choice questions. A passing score for this exam is set at 65%. Every year the Israel Medical Association publishes the exam itself, the average scores among the test takers, and the passing percentage. In 2022, the average score was 68% with 65% of participants passing the exam.

Large language models represent an exciting tool that can be utilized in many information gathering situations. Chat generative pre-trained transformer (ChatGPT) 3.5 is a free and easy-to-use language-based model that can be accessed as a tool to aid in test preparation [1]. Previous studies suggest that ChatGPT can pass medical exams in English [2]. It was unclear whether these results can be duplicated in Hebrew.

The aim of this study was to determine whether ChatGPT 3.5 could pass the Shlav *Aleph* exam in Hebrew.

Table 1. Results of ChatGPT 3.5 answers on the 2023 National Internal Medicine *Shlav Aleph* examination

Topic	Correct answers	Incorrect answers	Total questions	% Correct
Biostatistics	1	0	1	100.00%
Cardiology	4	8	12	33.33%
Endocrinology	5	9	14	35.71%
Gastroenterology	4	5	9	44.44%
General medicine	13	21	34	38.24%
Hematology/ oncology	4	6	10	40.00%
Infectious disease	10	16	26	38.46%
Pulmonary	3	6	9	33.33%
Renal	1	5	6	16.67%
Rheumatology	5	7	12	41.67%

PATIENTS AND METHODS

This study was exempt from institutional review board approval as no patient data was acquired. The 2023 *Shlav Aleph* examination questions were obtained from the Israel Medical Association (IMA) website [3]. ChatGPT was prompted with the questions as well as the multiple-choice answers in Hebrew. Questions were input individually, and answers were recorded.

Only questions that were strictly textual were recorded for the analysis. All multiple-choice questions associated with an image were excluded. ChatGPT's answers were then compared with the official answer key provided by the IMA. The IMA website provides the official answer key before and after the national resident appeal.

Figure 1. ChatGPT's incorrect answers shown in orange

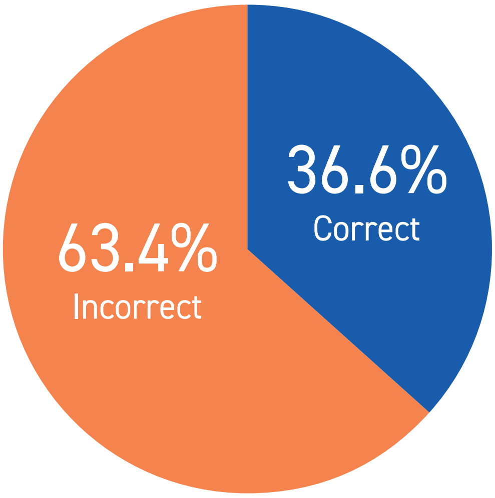
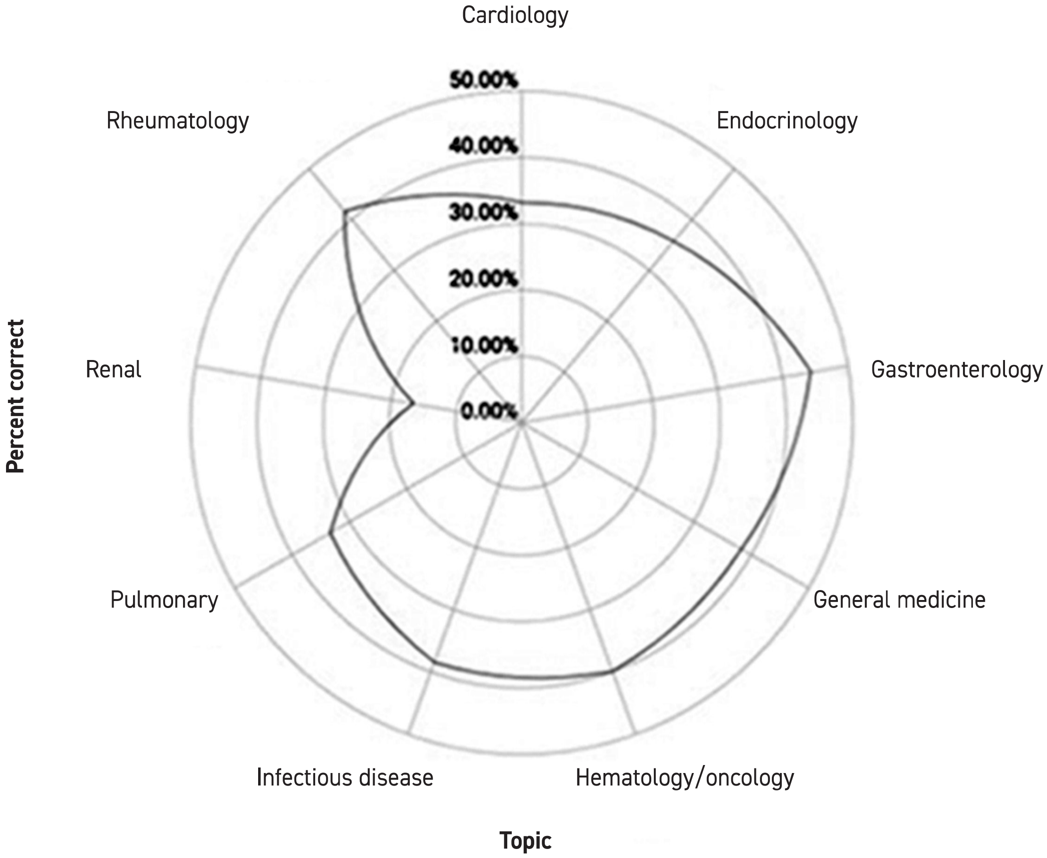


Figure 2. Spider Chart displaying that ChatGPT's performance was consistently inaccurate across various internal medicine subtopics



To ensure accuracy ChatGPT's answers were compared after the appeal. If after the appeal two answer choices were deemed correct, if ChatGPT chose either answer, it was marked as correct. Of 150 questions, 17 were excluded because they were associated with an image; 133 questions were available for analysis.

The 133 questions were analyzed and further categorized into topics to assess whether ChatGPT performed better in specific subject areas.

RESULTS

The results of our analysis showed that ChatGPT 3.5 answered 50/133 questions (36.6%) correctly. The AI algorithm performance was similar across all topics except questions in nephrology, where it performed poorly, and in biostatistics, with only one question in the exam, which the algorithm solved correctly [Table 1, Figure 1, Figure 2].

DISCUSSION

Artificial intelligence has many implications for healthcare. One of its applications is obtaining accurate medical information. This task has been tested with ChatGPT on various medical licensing exams, including the USMLE and the U.S. neurosurgical, radiology, and neurology board exams [4]. In these cases, ChatGPT 3.5 passed the exam. However, in other languages ChatGPT did not perform as well [5,6]. In the internal medicine *Shlav Aleph* in Hebrew, ChatGPT did not receive a passing grade. It performed poorly and much lower than the national average. In fact, it performed so poorly that it scored only slightly higher than if it were to answer the questions at random.

Several factors may have influenced its performance in Hebrew in contrast to English. First, the model has extensive training data in English, potentially surpassing that of Hebrew. This greater exposure to English medical terminology and concepts may have contributed to better performance in English.

Second, the complexity of translation poses challenges because the model, while multilingual, may encounter intricacies in Hebrew medical terminology not present in English. In addition, board exams assess not only medical knowledge but also country-specific guidelines and practices. Last, Hebrew's distinct language structure may pose challenges, with some concepts not translating perfectly.

CONCLUSIONS

While ChatGPT has shown promise in English medical exams, its adaptation to Hebrew presents several challenges that warrant further investigation.

Correspondence

Dr. D.J. Ozeri
Division of Rheumatology, Sheba Medical Center, Tel Hashomer 52621, Israel
Email: davidjoshua.ozeri@sheba.health.gov.il

References

1. Ozdemir S. Quick Start Guide to large language models: strategies and best practices for using ChatGPT and other LLMs. Addison-Wesley Data and Analytics Series, Pearson Education (US), 2023. ISBN 9780138199197.
2. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023; 9: e45312.
3. Scientific Council Israel Medical Association. 12 June 2023. Internal Medicine Shlav Aleph Exam. [Available from ima.org.il. https://ima-contentfiles.s3.amazonaws.com/examInternalMedicine01062023.pdf].
4. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. *medRxiv* [Preprint]. 2023 Feb 7: 2023.02.02.23285399.
5. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. *Sci Rep* 2023; 13: 20512.
6. Kılıç ME. AI in medical education: a comparative analysis of GPT-4 and GPT-3.5 on Turkish Medical Specialization Exam Performance. *medRxiv* [Preprint] 2023 Feb 7: 2023.02.02.23285399.

What loneliness is more lonely than distrust?

George Eliot (Mary Ann Evans) (1819–1880), English novelist, poet, journalist, translator

He who, when called upon to speak a disagreeable truth, tells it boldly and has done, is both bolder and milder than he who nibbles in a low voice and never ceases nibbling.

Johann Kaspar Lavater (1741–1801), Swiss poet, writer, philosopher, physiognomist and theologian