Examining ChatGPT Performance on USMLE Sample Items and Implications for Assessment

Victoria Yaneva, PhD, Peter Baldwin, EdD, Daniel P. Jurich, PhD, Kimberly Swygert, PhD, and Brian E. Clauser, EdD

Abstract

Purpose

In late 2022 and early 2023, reports that ChatGPT could pass the United States Medical Licensing Examination (USMLE) generated considerable excitement, and media response suggested ChatGPT has credible medical knowledge. This report analyzes the extent to which an artificial intelligence (AI) agent's performance on these sample items can generalize to performance on an actual USMLE examination and an illustration is given using ChatGPT.

Method

As with earlier investigations, analyses were based on publicly available USMLE sample items. Each item was submitted to ChatGPT (version 3.5) 3 times to evaluate stability. Responses were scored following rules that match operational practice, and a preliminary analysis explored the characteristics of items that ChatGPT answered correctly. The study was conducted between February and March 2023.

Results

For the full sample of items, ChatGPT scored above 60% correct except for one replication for Step 3. Response success varied across replications for 76 items (20%). There was a modest correspondence with item difficulty wherein ChatGPT was more likely to respond correctly to items found easier by examinees. ChatGPT performed significantly worse ($P < .001$) on items relating to practice-based learning.

Conclusions

Achieving 60% accuracy is an approximate indicator of meeting the passing standard, requiring statistical adjustments for comparison. Hence, this assessment can only suggest consistency with the passing standards for Steps 1 and 2 Clinical Knowledge, with further limitations in extrapolating this inference to Step 3. These limitations are due to variances in item difficulty and exclusion of the simulation component of Step 3 from the evaluation—limitations that would apply to any AI system evaluated on the Step 3 sample items. It is crucial to note that responses from large language models exhibit notable variations when faced with repeated inquiries, underscoring the need for expert validation to ensure their utility as a learning tool.

During recent months, a great deal of attention has been given to ChatGPT[1] and other advanced applications of artificial intelligence (AI). In medical contexts, these applications include medical note taking, consultation, diagnosis, and education—and research into these and other areas is growing (e.g., the *New England Journal of Medicine* has announced a new journal focusing on these applications to begin publication in 2024).[2–4] One measure of the capabilities of AI systems is their success at answering medical test questions, and several claims have been made about these systems passing the United States Medical Licensing Examination (USMLE). A January 2023 headline in *MedPage Today* announced, "AI passes U.S. Medical Licensing Exam—two papers show that large language models, including ChatGPT, can pass the USMLE."[5] *Medscape* followed with "AI Bot ChatGPT Passes U.S. Medical Licensing Exams without cramming—unlike students."[6] The careful reader of the original publications upon which these subsequent write-ups are based will notice that the claims made by journalists are not always well aligned with those results reported by researchers—and even that some of the conclusions stated in the original research are inconsistent with the reported data analysis. Furthermore, the findings reported in the original research are inconsistent across studies. In this article, we attempt to clarify what has been asserted in previous research. We follow this by replicating and extending previous work, and then examine the extent to which previous findings (as well as our own) can be meaningfully compared to the passing scores for the 3 USMLE Step exams. Finally, we reflect on the implications of these collective findings for the future of medical education and assessment.

There are at least 4 publications that have been referenced as evidence with respect to claims about AI systems passing different Steps of the USMLE.[7–10] It is crucial to note that none of these AI systems have interacted with an actual USMLE examination—that material is secure and not available to external researchers. The claims about performance are based on responses to other sets of items that at best approximate USMLE examinations. For some studies, the test material comprises large sets of multiple-choice questions (MCQs) assessing medical knowledge that have little if any actual USMLE content.[9,10] In other instances, the

Please see the end of this article for information about the authors.

Correspondence should be addressed to Victoria Yaneva, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104; email: vyaneva@nbme.org.

Acad Med. 2024;99:192–197.
First published online November 7, 2023
doi: 10.1097/ACM.0000000000005549

analyses have been based on sample items made available by the USMLE program.[7,8]

The study by Singhal and colleagues[10] applies a relatively new AI system (Flan-PaLM) to different data sets including one referred to as MedQA. While the authors state that MedQA is a "dataset comprising USMLE questions," elsewhere, they describe it more accurately as principally containing only USMLE "style" questions. MedQA does include publicly available USMLE sample items, but the vast majority of items were scraped from other sites offering test preparation for the USMLE rather than USMLE items.[11] Nonetheless, Singhal and colleagues reported a 67.6% accuracy level for Flan-PaLM, which is impressive regardless of the source of the questions.

Likewise, research by Liévin and colleagues[9] also refers to USMLE items but appears to have used the same MedQA data set. The performance they report is more than 6 percentage points below that for Flan-PaLM; however, comparisons across studies can be difficult even when researchers use the same data set. Most researchers in the aforementioned studies[7–10] excluded items that include nontext components (e.g., graphs, images), while others additionally excluded items for which the AI system did not provide an unambiguous answer. Some have also limited their sample to questions with 4 answer options.

In addition to the Singhal and Liévin groups, other researchers have tested systems using the MedQA data set.[12,13] Those systems do not appear to have performed as well as those described by Singhal and colleagues[10] or Liévin and colleagues[9]—but taken together, these studies show substantial progress in the accuracy with which AI systems can answer MCQs with medical content. Nonetheless, they do not provide a basis for comparing that level of performance to the level required to pass USMLE examinations, because the difficulty of the MedQA items may not be similar to that for items from a USMLE test form.

In contrast, there are 2 studies based on publicly available items that were previously used on USMLE.[7,8] We give these studies particular attention. Research by Kung and colleagues[8] began with a sample of 376 publicly available test questions downloaded from the USMLE website. After excluding items with nontext components, 305 (81%) items remained (93 Step 1 items, 99 Step 2 Clinical Knowledge (CK) items, and 113 Step 3 items). Each of these 305 items was presented in 3 ways: as an open-ended item, an MCQ, and an MCQ with a justification requested from ChatGPT, version 3.5 (powered by OpenAI Global LLC, San Francisco, CA). Because only the second of these approaches is directly relevant to the USMLE (and can be scored objectively), we focus on those results. Kung and colleagues present 2 sets of results: one with "indeterminate" responses censored and one with these responses included, stating that the censored/included accuracy for ChatGPT "for USMLE Steps 1, 2 CK, and 3 was 55.1%/36.1%, 59.1%/56.9%, and 60.9%/54.9%, respectively."

Gilson and colleagues[7] also examined the performance of ChatGPT 3.5 on USMLE sample items, although they limited their evaluation to Step 1 and Step 2 CK items. The authors state that before excluding items with a nontext component, they had a sample of 120 items for each Step examination. After excluding these items, they report findings for 87 Step 1 items (73%) and 102 Step 2 CK items (85%) with an accuracy level of 64.4% and 57.8%, respectively.

Although it appears that the studies by the Kung and Gilson groups accessed the same publicly available USMLE items, their studies were based on different numbers of items and findings were substantially different. Kung and colleagues[8] found that ChatGPT provided an "indeterminate" response for approximately one-third of the Step 1 items in the sample and provided a correct response to only 36.1% of the 93 items presented. In contrast, Gilson and colleagues[7] reported a correct response to 64.4% of the 87 items presented. Findings from both were much more similar for the Step 2 CK items.

The inconsistency between studies creates uncertainty about the findings, which is exacerbated by the claims made about how ChatGPT's performance compares to the passing score for the USMLE. Having stated that the USMLE passing score is approximately 60%, Kung and colleagues[8] then claim: "ChatGPT is now comfortably within the passing range." However, when USMLE examinees fail to respond to a question, it is scored as incorrect. Applying that scoring rule in this study, ChatGPT scored at 36.1%, 56.9%, and 54.9% on Steps 1, 2 CK, and 3, respectively. None of these values is above the approximate passing standard. (Typically, examinees must answer approximately 60% of items correctly to pass the Step examinations. This value is an approximation because the score scale and passing standard are not based on percent correct.)

The purpose of this study is therefore to provide clarity on the extent to which an AI agent's performance on these sample items could be indicative of its performance on the USMLE. To that end, we present a case study using ChatGPT (version 3.5). We begin by replicating these previous studies by presenting the publicly available USMLE items to ChatGPT; however, unlike previous studies, we present each item to ChatGPT 3 times, which allows us to examine the consistency of ChatGPT's performance. We also separately evaluate the text-only items (examined in previous publications[7,8]) as well as the items that included nontext elements (e.g., images). We conduct preliminary analyses to identify the characteristics of items that ChatGPT answered correctly and compare its performance with proportion correct responses from USMLE examinees. Additionally, we addressed the extent to which the publicly available USMLE items can be considered equivalent to a full examination and the limitations of generalizing the passing score to other item sets.

Method

As the Kung[8] and Gilson[7] research groups did, we used the USMLE items available from the USMLE website in early 2023.[14] This sample comprised 120 Step 1 items, 120 Step 2 CK items, and 137 Step 3 items. (Three items appeared in both the Step 2 CK and Step 3 sample item sets.) These sets of sample items are substantially smaller than operational USMLE test forms (Step 1 forms have approximately 280 items; Step 2, approximately 318 items; and Step 3, approximately 503 items). Moreover, operational forms are built to conform to

a complex set of content and statistical specifications that (among other objectives) minimize differences in difficulty between forms. In contrast, the publicly available sample USMLE item sets are built to condensed versions of the content blueprint, with less rigorous control of overall difficulty. Consequently, while the Step 1 and Step 2 CK sample forms were generally comparable to the operational USMLE form difficulties, the Step 3 items were notably less difficult.

As Kung and colleagues[8] noted, the sample items were made public in 2022, after ChatGPT 3.5 was trained, providing strong evidence that ChatGPT could not have "memorized" the answers during training. While ChatGPT accepts a wide range of special characters and symbols, approximately 14% of the sample items contain nontext features (e.g., images, graphs), which ChatGPT cannot interpret. Nevertheless, for completeness, we presented the text for every item to ChatGPT. We separately analyzed the accuracy for items that were entirely text based and items that originally included a nontext component. A new session was initiated for each item. We collected the responses from ChatGPT 3.5 between February 20, 2023, and March 19, 2023.

Replications

Consistent with the approach used in previous studies, we presented each item to ChatGPT verbatim as it would be presented to examinees, including presenting each option on a separate line and without providing additional instructions. We repeated this process 3 times for each item (replications) to evaluate the intra-item consistency of the responses.

Scoring ChatGPT responses

Each of the 3 responses to each item was then independently scored by 2 raters using a rubric specifying that a response was to be scored as correct only if ChatGPT identified the keyed answer as the only correct response. We scored all other variations of responses as incorrect, including occasions when an incorrect option was indicated, no answer was indicated, an answer was indicated that was not among the options, or more than one option was indicated as correct. This ensured that the scoring matched operational practice, where, for example, selecting 2 options is not permitted. Scoring disagreements between the 2 raters were reviewed by the first author (V.Y.) who made the final scoring decision.

In addition to evaluating the overall performance for ChatGPT, we conducted preliminary analyses to identify the characteristics of items that ChatGPT answered correctly. Because all items had, at one time, appeared on actual USMLE tests, we first compared ChatGPT's performance with the proportion of correct responses (*P* values) calculated using examinee responses to these items from first-time examinees from Liaison Committee on Medical Education (LCME) accredited medical schools. This was accomplished by examining the correlations between these *P* values and the 0/1 (incorrect/correct) scores from ChatGPT to determine the extent to which items that were more difficult for examinees were also more difficult for ChatGPT. We also considered the variability of ChatGPT's performance across one of the major USMLE content coding schemes, physician task competency, which is a framework for assigning each item to a distinct physician competency such as foundational science or diagnosis.[15] To evaluate the statistical significance of differences of performance across content, we tested each content area separately using a permutation test (100,000 permutations, 2-tailed). A Holm–Bonferroni correction was made for multiple comparisons to reduce Type 1 error. A permutation test is a form of proof by contradiction: the proposition that ChatGPT's success is affected by the task membership of an item is first assumed to be false, and then, if a contradiction arises (in this case, an observed difference that would be highly unlikely were the proposition false), this is interpreted as evidence that the proposition is true. In other words, if a highly unlikely difference is observed, this is evidence that ChatGPT performs differently on this content.

Results

Table 1 presents the percentage of items answered correctly by ChatGPT on the sample material for each of the 3 Steps' sample item sets across the 3 separate replications. Results are provided separately for the full sample of items, the items with text only, and the items with a nontext component.

Based on the full sample of items, ChatGPT scored above 60% correct in all cases except for one Step 3 replication, with a higher percentage of correct responses for text-only items.

Variability in the percentage of correct responses across the 3 replications is also shown in Table 1. An example of this variation is reported in Chart 1, which shows both a correct and incorrect response that ChatGPT produced for a single item. Such inconsistencies were observed for 76 (20%) of the 377 sample items (26 items for Step 1, 23 for Step 2 CK, and 27 for Step 3).

The relationship between item difficulty (measured by *P* values based on examinee responses collected when these items were used on operational Step examinations) and the responses from ChatGPT showed a modest correspondence, at best. The mean correlation across the 3 replications was 0.22 for Step 1, 0.23 for Step 2 CK, and 0.03 for Step 3 based on the full item sample. These correlations are similar for the text-only items (0.15, 0.20, and 0.07, respectively). Although some of the correlations for individual replications were significant ($P < .05$), none suggested a strong relationship.

Table 2 presents the proportion of items answered correctly by ChatGPT within each physician task competency assessed on the sample material. The observed difference between performance within and outside each task is also reported, along with the probability that each difference (or greater) would arise due to chance were the true difference zero. ChatGPT performed significantly worse ($P < .001$) on items relating to practice-based learning (which cover the topics of biostatistics, epidemiology, research ethics, and regulatory issues) than it did on other items. Items belonging to different content areas were not necessarily equally difficult; the data reported in Table 2 do not account for this possibility.

Discussion

Although it is impossible to make a precise and definitive statement about "passing" based on our findings

Table 1

Percentage of USMLE Sample Items Answered Correctly by ChatGPT Across 3 Replications, From a Study of ChatGPT Response Accuracy and Characteristics, 2023[a]

Examination	No. of items	Replication 1	Replication 2	Replication 3
Complete sample	377			
Step 1	120	60.83	64.17	64.17
Step 2 CK	120	70.83	67.50	71.67
Step 3	137	59.12	58.39	63.50
Text-only items	325			
Step 1	93	66.67	69.89	67.74
Step 2 CK	108	70.37	67.59	72.22
Step 3	124	60.48	59.68	65.32
Items with nontext component	52			
Step 1	27	40.74	44.44	51.85
Step 2 CK	12	75.00	66.67	66.67
Step 3	13	46.15	46.15	46.15

Abbreviations: USMLE, United States Medical Licensing Examination; CK, clinical knowledge.
[a]ChatGPT version 3.5 (powered by OpenAI Global LLC, San Francisco, CA).

(see Table 1), taken on average across replications, ChatGPT's performance appears consistent with "passing" for Step 1 and Step 2 CK. Given that the items in the Step 3 sample were easier to answer correctly than those on a typical Step 3 test form, ChatGPT's performance on the Step 3 sample item set would likely translate to a score below 60% correct on the operational exam for at least 2 of the 3 replications. Moreover, an operational Step 3 exam contains an interactive computer-based simulation component that contributes to the overall score, which we did not assess in this study. Note that while these findings are generally consistent with those presented in earlier publications,[7,8] they reflect uniformly higher performance for ChatGPT on the text-only items than was reported by Kung and colleagues[8] or Gilson and colleagues.[7]

The data reported in Table 1 are noteworthy in that they include ChatGPT's performance on all items in the sample set, including those that contained nontext elements. Perhaps not surprisingly, performance was lower on these items because ChatGPT can only interpret text. Nevertheless, ChatGPT's performance on items that originally included a nontext component far exceeded what would be expected by chance alone.

ChatGPT performed near or above the level associated with passing on the

Table 2

Proportion Correct for ChatGPT Responses to USMLE Sample Items Within USMLE Physician Tasks/Competencies, From a Study of ChatGPT Response Accuracy and Characteristics, 2023[a]

Item type	Physician task	No. of items		Proportion correct		Difference in proportion correct	Probability of observed difference[b]
		Within this physician task	Not within this physician task	Within this physician task	Not within this physician task		
Includes items with nontext elements (n = 377)	Communication[c]	32	345	.76	.63	.13	.06
	Foundational science	89	288	.58	.66	–.08	.06
	Diagnosis	125	252	.69	.62	.08	.06
	Management	106	271	.68	.63	.05	.18
	Practice-based learning	25	352	.32	.67	–.35	< .001
Excludes items with nontext elements (n = 325)	Communication[c]	31	294	.75	.65	.10	.12
	Foundational science	68	257	.65	.67	–.02	.36
	Diagnosis	107	218	.70	.64	.06	.12
	Management	95	230	.68	.66	.03	.28
	Practice-based learning	24	301	.33	.69	–.36	< .001

Abbreviation: USMLE, United States Medical Licensing Examination.
[a]ChatGPT version 3.5 (powered by OpenAI Global LLC, San Francisco, CA).
[b]*P* value significant at alpha = .05 (2-tailed). A correction for multiple comparisons was made using the Holm–Bonferroni method.
[c]The full name of the communication category is "communication/professionalism/systems-based practice and patient safety."

multiple-choice component for Steps 1 and 2 CK. Inferences for passing Step 3 are limited, both because the exam contains a simulation component unassessed in this study and because of the comparative easiness of the sample items noted earlier. That said, it is worth emphasizing that ChatGPT performed well below a typical USMLE examinee. In contrast, average scores for first-time examinees from LCME-accredited schools are approximately 2 standard deviations above the passing standard.[16] Moreover, although AI systems will certainly improve over time, there is an important distinction between performing well on these multiple-choice questions and being licensed to practice medicine. In addition to passing the USMLE sequence, a physician must successfully complete medical school and residency, which requires demonstrating a range of competencies not assessed by USMLE test material but critical to the provision of safe and effective patient care.

Because much of the attention given to ChatGPT has focused on performance relative to the USMLE passing standards, it is useful to consider the limitations of such interpretations. Answering 60% of items correctly is an approximation of the passing standard. Additionally, multiple forms of any Step examination can only be compared to a common cut score because they have been built to the same statistical and content specifications and have been statistically adjusted (equated) to place scores from different forms on the same scale.[17] As noted, the publicly available study items were assembled with less rigorous constraints, introducing differences in content representation and difficulty compared to an operational test form. As such, any comparisons made based on publicly available USMLE items will be approximate and, as appears to be the case for Step 3, may be an imprecise approximation. This limitation is a particularly important consideration when interpreting results based solely on text-based items. Materials representing evidence-based medicine, biostatistics, radiology, dermatology, and cardiology, which typically incorporate a higher proportion of nontext elements, will likely be underrepresented when items with nontext components are excluded.

Our findings also raise questions in several areas related to medical education and assessment. Others have suggested that AI will have an important role in these areas going forward[2,3]—which seems like a safe prediction. At present, however, there are limitations to the usefulness of ChatGPT for medical students. ChatGPT appears equally confident whether or not its answer is correct. Promoting the use of these tools as learning aides for medical students should be avoided without first emphasizing the need for expert review of the output. Whether AI systems can be useful aids to experts writing response rationales remains an open question that needs empirical investigation. For the time being, learning and assessment will be better supported by materials that have been more rigorously vetted.

Another issue worth examining is what these findings suggest about the cognitive processes required for examinees to respond to USMLE items. Although ChatGPT claims that it "is not capable of reasoning in the same way that humans do," as per its own response, many of its responses closely resemble human responses, despite being generated probabilistically based on word cooccurrences in a large corpus of human-generated text—raising the question of whether this form of probabilistic prediction is (or should be) included in a definition of "clinical reasoning." Yet such a question presupposes overlap between ChatGPT's response processes and examinees' response processes, which has not been demonstrated.

In a survey overview of automated question-answering systems, Rogers and colleagues[18] write, "It is increasingly clear that humans and machines do not necessarily find the same things difficult, which complicates direct comparisons of their performance." This view is consistent with the modest, and frequently nonsignificant, correlations we observed between the difficulty of items for examinees and the performance of ChatGPT on those items.

The question of the cognitive processes required to answer these questions is also brought into focus when we consider the items with images and other nontext components. Previous researchers have removed these items from their study sample—presumably under the assumption that ChatGPT would be unable to respond. This is clearly not the case, however, as we found that ChatGPT provided detailed descriptions of "imagined" graphs in its responses to items with nontext components, even though no graphs were included in the input. Without additional study in this area, it is not possible to disentangle whether ChatGPT was able to answer these questions because the image is not required to arrive at the correct answer, or because it was able to draw parallels between the text and image descriptions from its training data. An interesting avenue for future research would be to investigate whether examinees would be as successful in answering these items without access to the nontext component.

What, then, are the implications of AI systems successfully answering test items? It seems clear that recent advances represent a trend that is likely to continue. Although this general trend has substantial significance for the performance of the model itself (mainly related to how useful it may be for other tasks from that domain), the implications are more modest regarding education and assessment. If improved large language models were to answer more items correctly, educators and patients would still want physicians to be familiar with foundational concepts in medicine and able to use that knowledge to build advanced skills and experience. That said, if the question is whether we should embrace innovations with AI that physicians can use to improve the quality of care, our answer is a most enthusiastic yes. Any prediction of how this may happen is premature at this stage, but it is safe to assume that as the use of AI in medical practice continues to grow, its reach will extend to how we assess medical knowledge and skills.

In conclusion, it seems appropriate that we comment on how advances in AI will impact the responsibility that assessment organizations and examinees have to the public. Medical licensing examinations are explicitly intended to protect the health of the public. Both the form and content of these tests are intended to support inferences about specific proficiencies required for the safe and effective practice of medicine.[19] When new technologies change the way medicine is practiced or the way medical students learn, testing organizations must

reevaluate the alignment between the form and content of their tests and the function of those tests in protecting the public. This reevaluation is an essential and ongoing part of a testing organization's responsibility to evaluate the validity of the inferences and uses that are made based on the test scores they produce. Studies such as this represent a first step in that process of reevaluation.

Funding/Support: All work on the research and manuscript was completed while the authors were employees of the National Board of Medical Examiners.

Other disclosures: None reported.

Ethical approval: This research has been determined by the American Institutes of Research as not meeting the federal definition of research with human subjects and is therefore exempt from IRB review and oversight (Project No. EX00398).

Previous presentations: Annual meeting of the National Council on Measurement in Education, April 14, 2023, Chicago, IL.

Data sharing: All data are from the National Board of Medical Examiners, which approved its use prior to the manuscript's submission.

V. Yaneva is manager, Natural Language Processing Research, Office of Research Strategy, National Board of Medical Examiners, Philadelphia, Pennsylvania.

P. Baldwin is principal measurement scientist, Office of Research Strategy, National Board of Medical Examiners, Philadelphia, Pennsylvania.

D.P. Jurich is associate vice president, United States Medical Licensing Examination, National Board of Medical Examiners, Philadelphia, Pennsylvania.

K. Swygert is director, Test Development Innovations, National Board of Medical Examiners, Philadelphia, Pennsylvania.

B.E. Clauser is distinguished research scientist, Office of Research Strategy, National Board of Medical Examiners, Philadelphia, Pennsylvania.

References

1 OpenAI. GPT-4 Technical Report. arXiv. Published March 20, 2023. Accessed September 25, 2023. https://arxiv.org/abs/2303.08774.
2 Lee P, Bubeck S, Petro P. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388:1233–1239.
3 Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine. N Engl J Med. 2023;388:1201–1208.
4 Beam AL, Drazen JM, Kohane IS, Leong T-Y, Manrai AK, Rubin EJ. Artificial intelligence in medicine. N Engl J Med. 2023;388:1220–1221.
5 DePeau-Wilson M. AI passes U.S. medical licensing exam—two papers show that large language models, including ChatGPT, can pass the USMLE. MedPage Today. Published January 19, 2023. Accessed September 25, 2023. https://www.medpagetoday.com/special-reports/exclusives/102705.
6 Ault A. AI bot ChatGPT passes US medical licensing exams without cramming—unlike students. Medscape. Published January 26, 2023. Accessed September 25, 2023. https://www.medscape.com/viewarticle/987549.
7 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.
8 Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198.
9 Liévin V, Hother CE, Winther O. Can large language models reason about medical questions? arXiv. Published January 24, 2023. Accessed September 25, 2023. https://arxiv.org/pdf/2207.08143.pdf.
10 Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. arXiv. Published December 26, 2022. Accessed September 25, 2023. https://arxiv.org/pdf/2212.13138.pdf.
11 Jin D, Pan E, Oufattole N, Weng W-H, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci. 2021;11:6421.
12 Yasunaga M, Leskovec J, Liang P. LinkBERT: pretraining language models with document links. arXiv. Published March 29, 2022. Accessed September 25, 2023. https://arxiv.org/pdf/2203.15827.pdf.
13 Yasunaga M, Bosselut A, Ren H, et al. Deep bidirectional language knowledge graph pretraining. arXiv. https://arxiv.org/pdf/2210.09338.pdf. Published October 19, 2022. Accessed September 25, 2023.
14 United States Medical Licensing Examination. Prepare for Your Exam. Accessed September 25, 2023. https://www.usmle.org/prepare-your-exam.
15 United States Medical Licensing Examination. USMLE Physician Task/Competencies. Published 2020. Accessed September 25, 2023. https://www.usmle.org/sites/default/files/2021-08/USMLE_Physician_Tasks_Competencies.pdf.
16 USMLE Score Interpretation Guidelines. United States Medical Licensing Exam. Accessed October 17, 2023. https://www.usmle.org/sites/default/files/2022-05/USMLE%20Step%20Examination%20Score%20Interpretation%20Guidelines_5_24_22_0.pdf.
17 Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL, ed. Educational Measurement. 2nd ed. Washington, DC: American Council on Education; 1971: 508–600.
18 Rogers A, Gardner M, Augenstein I. QA dataset explosion: a taxonomy of NLP resources for question answering and reading comprehension. ACM Comput Surv. 2023;55:1–45.
19 Margolis MJ, Clauser BE, Swanson DB. Issues of validity and reliability for assessments in medical education. In Holmboe ES, Durning SJ, eds. Practical Guide to Evaluation of Clinical Competence. Philadelphia, PA: Elsevier; in press.