Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01650-x

Synthetic medical education in dermatology leveraging generative artificial intelligence

Check for updates

Arya S. Rao[1,2], John Kim[1,2], Andrew Mu[1,2], Cameron C. Young[1,2], Ezra Kalmowitz[1,2], Michael Senter-Zapata[1,3], David C. Whitehead[1,4], Lilit Garibyan[1,4], Adam B. Landman[1,3] & Marc D. Succi[1,2,4]

The advent of large language models (LLMs) represents an enormous opportunity to revolutionize medical education. Via "synthetic education," LLMs can be harnessed to generate novel content for medical education purposes, offering potentially unlimited resources for physicians in training. Utilizing OpenAI's GPT-4, we generated clinical vignettes and accompanying explanations for 20 skin and soft tissue diseases tested on the United States Medical Licensing Examination. Physician experts gave the vignettes high average scores on a Likert scale in scientific accuracy (4.45/5), comprehensiveness (4.3/5), and overall quality (4.28/5) and low scores for potential clinical harm (1.6/5) and demographic bias (1.52/5). A strong correlation ($r = 0.83$) was observed between comprehensiveness and overall quality. Vignettes did not incorporate significant demographic diversity. This study underscores the potential of LLMs in enhancing the scalability, accessibility, and customizability of dermatology education materials. Efforts to increase vignettes' demographic diversity should be incorporated to increase applicability to diverse populations.

In a rapidly evolving technological landscape, medical education stands to reap significant benefits. As new generations of medical students and residents enter training, they are faced with a wealth of technologies that can serve as adjuncts to their learning, and the medical system is adapting. In particular, large language models (LLMs), a type of machine learning model, have gained notoriety for their vast computational power in recent months. These models, trained on vast amounts of textual data from a variety of sources, are able to harness these collective insights to perform highly specialized tasks. Generalist models like the generative pre-transformer (GPT) model from OpenAI have shown impressive performance in the clinical realm even without explicit training, indicating that the use of LLMs in medicine is only just beginning[1–14].

LLMs offer unprecedented utility in medical education given their ability to synthesize novel output rapidly and efficiently. Despite widespread interest in applying LLMs toward medical education tasks[1], there is little research on how LLM-guided education initiatives may actually perform in practice. The generation of novel clinical vignettes is one transformative yet unstudied application of LLMs in this sector.

Clinical vignettes are an important cornerstone of modern medical education, contributing both to the bulk of United States Medical Licensing Examination (USMLE) questions and preclinical case-based teaching[15]. Vignettes contextualize medical knowledge into practical scenarios that assess diagnostic reasoning, management prioritization, and understanding of psychosocial contexts, emulating the multi-layered and nuanced practice of medicine for the learner[16]. Traditionally, these resources have been accessible through professional societies (e.g., Undergraduate Web-Based Interactive Self-Evaluation (uWISE, https://apgo.org/page/uwisev3-2), Standardized Tool for the Assessment of Radiology Students (STARS, https://www.acr.org/Lifelong-Learning-and-CME/Learning-Activities/Medical-Student-Activities/Radiology-Student-Assessment-Tool)), in-house materials written by faculty, or commercially available question banks like UWorld and AMBOSS. The generation of these vignettes is a labor-intensive task, requiring the extensive input of experienced physicians. Therefore, while these sources promise a degree of quality control, the accessibility and quantity of these materials are unequally distributed to medical students across the nation, depending on factors like institutional access and student socioeconomic status. In addition, there is increasing concern over repeated test questions on USMLE administrations, caused in part by the scarcity of vignettes.

[1]Harvard Medical School, Boston, MA, USA. [2]Medically Engineered Solutions in Healthcare Incubator, Innovation in Operations Research Center (MESH IO), Mass General Brigham, Boston, MA, USA. [3]Brigham and Women's Hospital, Boston, MA, USA. [4]Massachusetts General Hospital, Boston, MA, USA.
e-mail: msucci@mgh.harvard.edu

While inherently driven by visual evaluation, medical instruction on many skin and soft tissue pathologies is inseparable from the holistic clinical presentation contextualizing the disease process, and standardized exams like the USMLE rely heavily upon text-based vignettes for these pathologies. Furthermore, the vernacular surrounding the description of skin lesions is key to the diagnosis and treatment of cutaneous disease[17,18]. LLMs offer a distinct capability to expand the availability of text-based vignettes for common dermatologic conditions in medical education. Furthermore, current off-the-shelf LLMs like GPT offer flexibility in expanding on the clinical vignette beyond the initial output, evolving to meet the needs of the individual student as subsequent questions arise. In this study, we evaluate the feasibility of utilizing GPT 4.0, OpenAI's latest public-facing foundation model, to generate high-quality clinical vignettes for the purpose of medical education.

Results

Vignette details

In the 20 clinical vignettes generated, 15 patients were male and 5 were female. The median patient age was 25.0 years; ages were always provided in multiples of 5 or 2. Race was provided for 4 patients; 3 were Caucasian and 1 was African American [Table 1]. Generic names (Mr. Johnson, Mr. Smith) were provided for 3 patients; all other vignettes did not include names.

The average word count for model output was 332.68, with a standard deviation of 42.75 words. The clinical vignette portion of the responses had 145.79 words on average, with a standard deviation of 26.97. The average length of explanations was 184.89, with a standard deviation of 49.70. Explanations were generally longer than their corresponding vignettes; the average ratio between the length of the vignette and the length of the explanation was 0.85, with a standard deviation of 0.30.

Physician scoring

Mean ratings reflect high levels of alignment with scientific consensus (4.45, 95% CI: 4.28–4.62), comprehensiveness (4.3, 95% CI: 4.11–4.89), and overall quality (4.28, 95% CI: 4.10–4.47) [Fig. 1]. Ratings also show low suspicion of clinical harm (1.6, 95% CI: 1.38–1.81) and demographic bias (1.52, 95% CI: 1.31–1.72). Assessment of demographic bias included evaluating whether vignettes contained stereotypical or disproportionately skewed representations of patient populations. The consistently low ratings suggest that no strong patterns of bias were detected by physician raters. Additionally, errors related to factual inaccuracy, including potential AI hallucinations, were indirectly captured in ratings for scientific alignment and clinical harm.

The criteria were also evaluated for internal correlations via the calculation of Pearson correlation coefficients. Alignment with scientific consensus was moderately correlated with comprehensiveness and overall quality, with correlation coefficients $r = 0.67$ and $r = 0.68$, respectively [Fig. 2]. Comprehensiveness and overall quality were strongly correlated with a correlation coefficient of $r = 0.83$. The possibility of clinical harm and the possibility of demographic bias were weakly correlated, with a correlation coefficient of $r = 0.22$.

Discussion

Considering recent scrutiny of standardized medical examinations, it is more important than ever that there is an abundance of high-quality educational materials to be used for standardized assessments like the USMLE. However, creating new questions is resource-intensive, requiring experienced physicians to write clinical vignettes and multiple test administrations to evaluate the generalizability of question performance. Novel methods for developing numerous, unique clinical vignettes are necessary and sought after.

In this study, we present promising evidence towards the feasibility and effectiveness of using a large language model, GPT-4, as a source of synthetic medical education, offering the potential for accessible, customizable, and scalable educational resources. We provide the first analysis of GPT-4's utility for clinical vignette generation, demonstrating that GPT-4's inherent clinical knowledge extends to the creation of representative and accurate patient descriptions. We find that for diseases tested in the Skin & Soft Tissue section of the USMLE Step 2 CK exam, GPT-4 generated vignettes that were highly accurate, highlighting the potential of LLMs to design vignettes which could eventually be incorporated into standardized medical examinations.

Analysis of the generated vignettes revealed high ratings in alignment with scientific consensus, comprehensiveness, and overall quality, coupled with low ratings in potential for clinical harm and demographic bias. There was a high statistical correlation between vignette comprehensiveness and overall quality ($r = 0.83$), indicating the importance of thorough and detailed case presentations in medical education and highlighting the ability of LLMs to provide contextually relevant and complete scenarios for clinical reasoning.

The average length of vignettes was 145.79 ± 26.97 words. This is well within the scope of USMLE vignette length; examinees have on average 90 s to answer each question on the USMLE Step 1, Step 2 CK, and Step 3 exams. Vignettes were accompanied by longer explanations, showcasing the ability of LLMs to generate not just patient descriptions, but also useful didactic material.

While vignettes received overall low ratings from evaluators for the possibility of demographic bias (1.52, 95% CI: 1.31– 1.72), the limited variety in patient demographics, highlighted by predominantly male patients and limited racial diversity, suggests a need for more conscious efforts to include diverse patient representations. Specific inclusion of such efforts in prompt engineering and model training datasets is crucial in preparing students to serve as physicians in an inclusive healthcare environment. Our study did not account for patient diversity in LLM prompts; further, since a new chat session was started for each prompt, overall balance could not be controlled by the LLM. In addition, future iterations of this work should further investigate sources and manifestations of systemic bias in model output.

While our initial pilot shows that GPT-4–generated vignettes display high clinical accuracy as evaluated by expert raters, it is important to note that LLM hallucinations may produce inconsistencies when deployed at scale. Additionally, LLMs are trained on the entire breadth of content available on the internet, which may not represent the standard of care. This could result in inaccurate responses. Deployment and widespread adoption of these models necessitates careful screening; clinical experts may be employed, as in this study, as evaluators of vignettes prior to usage. Specific training data for diagnoses of interest based on expert-recommended content may also help to refine model output and could facilitate the development of custom models.

A key limitation of this study is the composition of our expert rater panel, which included only one dermatologist alongside two attending physicians from internal medicine and emergency medicine. While these non-dermatologist raters frequently diagnose and manage common skin conditions in their respective specialties, and necessarily are familiar with the standard presentations of those evaluated in this study via their presence on national board examinations, their expertise may not encompass the full spectrum of dermatologic disease. As a result, their assessments were likely most reliable for clear-cut, board-style presentations but may have been less sensitive to subtle diagnostic nuances. Future studies would benefit from a larger proportion of dermatologists to ensure a more specialized evaluation of AI-generated cases.

Overall, this work demonstrates that off-the-shelf LLMs like GPT-4 hold great potential to be used for clinical vignette generation for standardized examination and teaching purposes. Fit-for-purpose LLMs trained on more specific datasets may further enhance these capabilities. The high accuracy and efficiency of "synthetic education" are a promising solution to current limitations in traditional means for generating medical educational materials.

Methods

Vignette generation

Of the 89 conditions tested under the USMLE Content Outline's Skin & Subcutaneous Tissue subheading (https://www.usmle.org/sites/default/

files/2022-01/USMLE_Content_Outline_0.pdf), the following 20 were randomly selected for inclusion in this study: Scarlet Fever, Local Subcutaneous Reaction, Hyperhidrosis, Stomatitis, Acne Vulgaris, Staphylococcal Scalded Skin Syndrome, Ichthyosis, Impetigo, Stevens-Johnson Syndrome, Tinea Corporis, Cauliflower Ear, Dermatoses Caused by Plants, Lentigo, Frostbite, Melanoma, Herpes Zoster, Cellulitis, Folliculitis, Pigmented Nevi, and Keloids. GPT-4 (OpenAI) was used to generate a clinical vignette on each topic using the following prompt: "Generate a detailed clinical vignette for the purpose of medical education on the topic of [condition] at the level of a preclinical student. In addition, include a sentence at the end about why [condition] is the most likely diagnosis and the components of the vignette that support it. Also provide an explanation of why the others on the differential are less likely based on the vignette." Responses were generated on November 14, 2023. A new chat session was used to generate each vignette. All vignettes are available in the Supplementary Materials.

Table 1 | Demographic details from generated clinical vignettes

Condition	Age	Gender	Race
Scarlet Fever	6	Male	Not provided
Local Subcutaneous Reaction	25	Female	Not provided
Hyperhidrosis	25	Male	Not provided
Stomatitis	45	Female	Not provided
Acne Vulgaris	16	Male	Not provided
Staphylococcal Scalded Skin Syndrome	2	Male	Not provided
Ichthyosis	6	Male	Not provided
Impetigo	6	Male	Not provided
Stevens-Johnson Syndrome	16	Male	Not provided
Tinea Corporis	10	Male	Not provided
Cauliflower Ear	25	Male	Not provided
Dermatoses Caused by Plants	35	Female	Not provided
Lentigo	45	Female	Caucasian
Frostbite	45	Male	Not provided
Melanoma	65	Male	Caucasian
Herpes Zoster	65	Male	Not provided
Cellulitis	65	Male	Not provided
Folliculitis	25	Male	Not provided
Pigmented nevi	25	Male	Caucasian
Keloids	25	Female	African American

Evaluation criteria

Vignettes were assessed for alignment with scientific consensus, possibility of clinical harm, comprehensiveness, possibility of demographic bias, and overall quality, based on previous criteria used to assess LLMs[19]. Three practicing attending physicians commonly encountering the topics at hand at our institution (emergency medicine, dermatology, and internal medicine) rated the answers based on how well they fit the criteria using a Likert scale (coded 1–5 respectively: from very opposed to very aligned with scientific consensus, very low possibility of harm to very high possibility of harm, very non-comprehensive to very comprehensive, very low possibility of demographic bias to very high possibility of demographic bias, and very poor quality to very high quality). Given the subjective nature of the Likert scale, we utilized an ensemble scoring strategy (https://www.aconf.org/conf_160152.htm), averaging scores across the three evaluators; 95% confidence intervals represent the inherent uncertainty in this metric.

Statistical methods

Statistical analyses were conducted using Python (3.12.1, Python Software Foundation, Wilmington, DE) packages *pandas* and *numpy*. Visualization was completed using the *matplotlib* package. The Pearson correlation coefficient was computed to assess the degree of correlation between evaluation criteria. A p-value of less than 0.05 was considered statistically

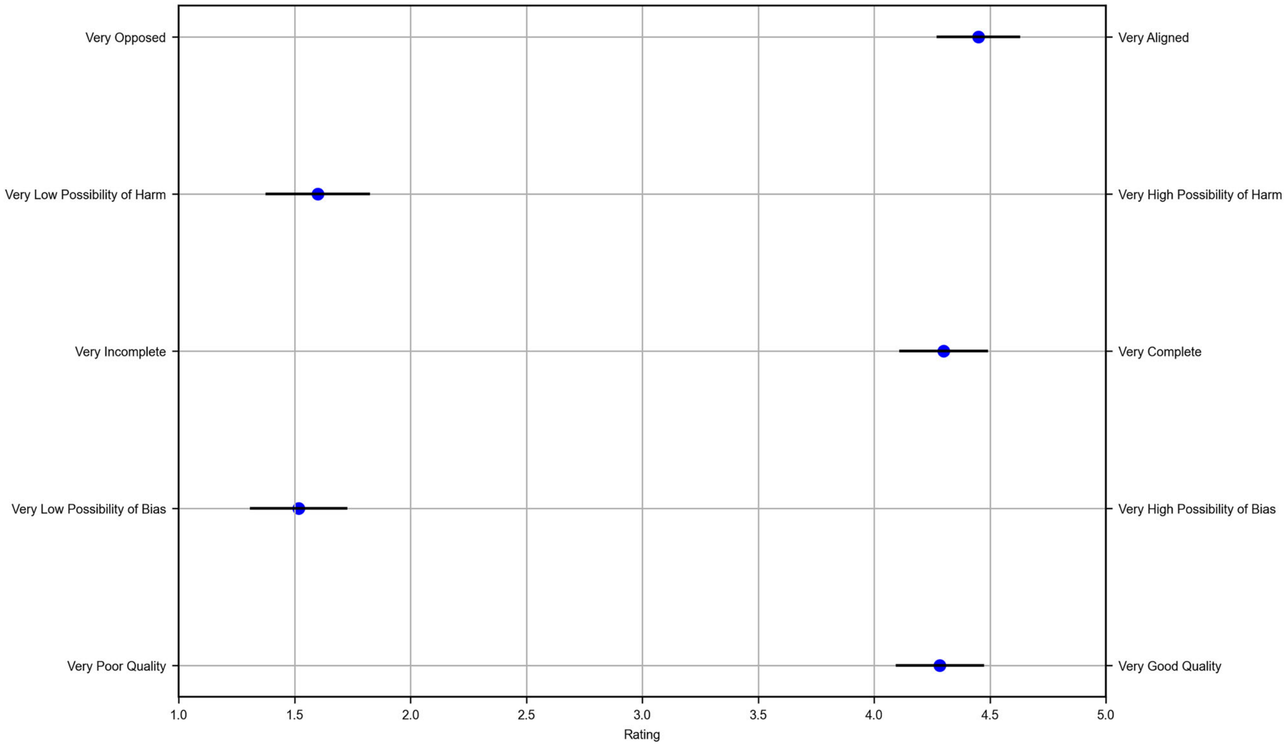

Fig. 1 | Clinical vignette ratings by three attending physicians. Average ratings by three attending physicians of the twenty clinical vignettes based on alignment with scientific consensus, possibility of clinical harm, comprehensiveness, possibility of demographic bias, and overall quality. 95% confidence intervals are included.

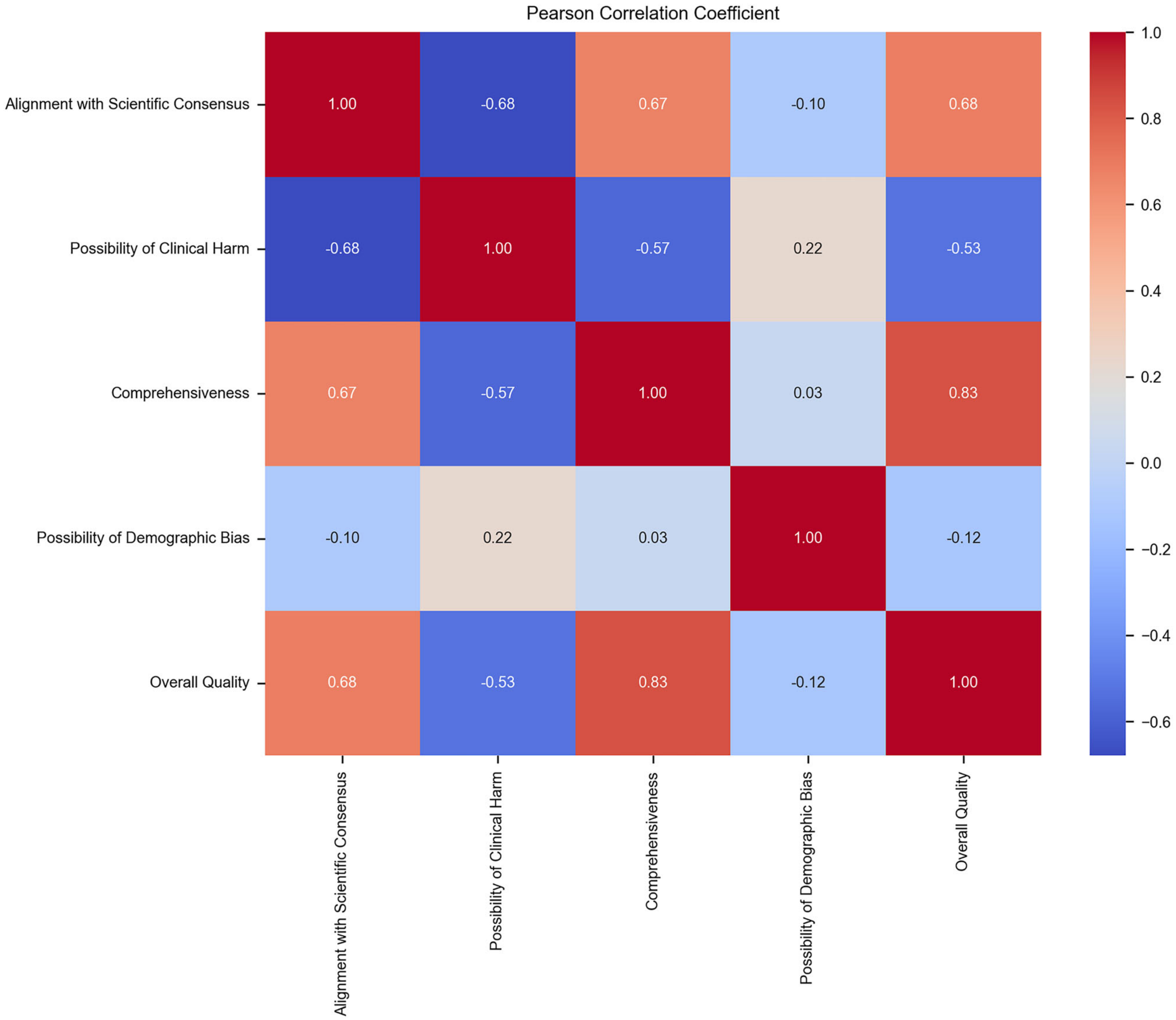

Fig. 2 | Correlations between evaluation criteria. Pearson correlation coefficients between the evaluation criteria of alignment with scientific consensus, possibility of clinical harm, comprehensiveness, possibility of demographic bias, and overall quality.

significant. This study received an IRB exemption as it did not involve patient data or human subjects.

Data availability

All data generated or analyzed during this study are included in this published article (and its Supplementary Materials). Additional details are available from the corresponding author upon reasonable request.

Received: 16 January 2025; Accepted: 18 April 2025;
Published online: 04 May 2025

References

1. Chang, B. S. Transformation of Undergraduate Medical Education in 2023. *J. Am. Med. Assoc.* **330**, 1521 (2023).
2. Rao, A. et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. *J. Am. Coll. Radiol.* **20**, 990–997 (2023).
3. Hirosawa, T. et al. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. *Int. J. Environ. Res. Public Health* **20**, 3378 (2023).
4. Rao, A. et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *J. Med. Internet Res.* **25**, e48659 (2023).
5. Dave, T., Athaluri, S. A. & Singh, S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front. Artif. Intell.* **6**, 1169595 (2023).
6. Koranteng, E. et al. Empathy and equity: key considerations for large language model adoption in health care. *JMIR Med. Educ.* **9**, e51199 (2023).
7. Young, C. C. et al. Diagnostic accuracy of a custom large language model on rare pediatric disease case reports. *Am. J. Med. Genet. A***197**, e63878 (2025).
8. Young, C. C. et al. Pilot study of large language models as an age-appropriate explanatory tool for chronic pediatric conditions. *medRxiv* https://doi.org/10.1101/2024.08.06.24311544 (2024).
9. Kim, J. et al. Risk stratification of potential drug interactions involving common over-the-counter medications and herbal supplements by a large language model. *J. Am. Pharm. Assoc*. https://doi.org/10.1016/j.japh.2024.102304 (2024).
10. Rao, A. et al. Proactive polypharmacy management using large language models: opportunities to enhance geriatric care. *J. Med. Syst.* **48**, 41 (2024).

11. Succi, M. D. & Rao, A. S. Beyond the AJR: towards large language models for radiology decision-making in the emergency department. *Am. J. Roentgenol*. https://doi.org/10.2214/AJR.24.32465 (2024).
12. Nguyen, D., Rao, A., Mazumder, A. & Succi, M. D. Exploring the accuracy of embedded ChatGPT-4 and ChatGPT-4o in generating BI-RADS scores: a pilot study in radiologic clinical support. *Clin. Imaging* **117**, 110335 (2025).
13. Rao, A. et al. A future of self-directed patient internet research: large language model-based tools versus standard search engines. *Ann. Biomed. Eng*. https://doi.org/10.1007/s10439-025-03701-6 (2025).
14. Rao, A. S. et al. A large language model-guided approach to the focused physical exam. *J. Med. Artif. Intell.* **8** (2024).
15. Besche, H. C., Schwartzstein, R. M., King, R. W., Hoenig, M. P. & Cockrill, B. A. *A Step-by-Step Guide to Case-Based Collaborative Learning (CBCL)*. (Springer International Publishing, Cham, 2022). https://doi.org/10.1007/978-3-031-14440-0.
16. Nendaz, M. R., Raetzo, M. A., Junod, A. F. & Vu, N. V. Teaching diagnostic skills: clinical vignettes or chief complaints?. *Adv. Health Sci. Educ.* **5**, 3–10 (2000).
17. Papier, A., Chalmers, R. J. G., Byrnes, J. A. & Goldsmith, L. A. Framework for improved communication: the Dermatology Lexicon Project. *J. Am. Acad. Dermatol.* **50**, 630–634 (2004).
18. Fisher, H. M. et al. DermO; an ontology for the description of dermatologic disease. *J. Biomed. Semant.* **7**, 38 (2016).
19. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).

Acknowledgements

The project described was supported in part by award Number T32GM144273 from the National Institute of General Medical Sciences, which played no role in study design, data collection, analysis, and interpretation of data, or the writing of this paper. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

Author contributions

A.S.R. and M.D.S. conceived of and designed this study. A.S.R., A.M., and C.Y. were responsible for data generation, acquisition, analysis, and interpretation. M.S.Z., D.C.W., and L.G. served as physician reviewers for the vignettes generated. A.S.R., A.M., C.Y., E.K., J.K., and A.B.L. drafted this paper. All authors participated in the critical revision of the paper for important intellectual content. M.D.S. supervised this work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01650-x.

Correspondence and requests for materials should be addressed to Marc D. Succi.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.