scientific reports

Check for updates

OPEN

Performance of vision-language models compared with 252 medical students on text-only and image-based dermatology examinations

Ozan Erdem[1]✉, Abdurrahim Yilmaz[2]✉, Ahmet Sait Şahin[1], Buğra Burç Dağtaş[3], Ece Gökyayla[4], Melek Aslan Kayıran[1], Vefa Aslı Erdemir[1] & Mehmet Salih Gürel[1]

Vision–language models (VLMs) are increasingly evaluated in medical education, yet their performance on visually intensive assessments remains incompletely understood. We compared four state-of-the-art VLMs, GPT-4o and GPT-5 (both accessed via ChatGPT), Gemini 2.5 Flash, and Gemini 3 Pro, with fifth-year medical students on ten consecutive dermatology clerkship examinations administered between September 2023 and January 2025. Examinations combined text-only questions (multiple-choice, multiple-select, and matching; 60% of total score) with image-based, structured open-ended questions (40%) spanning seven dermatologic sub-domains. Model outputs were evaluated using expert-validated answer keys and grading rubrics, with repeated runs to assess output variability. All VLMs significantly outperformed medical students on text-only examinations (mean scores > 95 vs. 84.9; $p < 0.001$), showing minimal sensitivity to exam difficulty. In contrast, image-based performance was heterogeneous: Gemini 3 Pro and GPT-5 achieved higher scores than students, whereas students significantly outperformed GPT-4o and Gemini 2.5 Flash. Medical students showed the smallest difference between text-only and image-based scores within this examination setting. Among VLMs, sub-domain analyses suggested that strong visual description and diagnostic performance did not uniformly translate into etiological, differential diagnostic, or treatment-related reasoning. Within the constraints of this single-center, examination-based benchmark, these findings indicate that VLMs demonstrate strong text-based dermatologic knowledge, whereas multimodal performance remains uneven and model-dependent. These results should not be interpreted as evidence of standalone clinical competence, but rather as support for further controlled evaluation of VLMs as complementary, human-supervised tools in dermatology education.

Keywords Vision-language models, Medical education, Dermatology, Medical students, Medical examinations, Large language models

Artificial intelligence (AI) is increasingly influencing diagnostic workflows and educational methodologies in medicine[1–4]. This evolution has been driven by rapid advances in large language models (LLMs) and, more recently, vision–language models (VLMs), which are designed to process and synthesize complex biomedical data[5–10]. Dermatology, an inherently visual specialty, represents a particularly relevant setting for this multimodal development. Because dermatologic diagnosis relies on integrating visual cues with clinical context to capture fine pathophysiological nuances, the field provides a useful testbed for evaluating whether VLMs can approximate the visual interpretation and integrative reasoning required in clinical practice[11–13].

The current literature includes numerous evaluations of the performance of LLMs and VLMs on national medical examinations and specialty board assessments, in which models have been reported to reach or exceed human passing thresholds[14–18]. Comparable findings have also been reported in non-English national medical examinations, indicating that the performance of LLMs and VLMs is not confined to English-language

[1]Department of Dermatology and Venereology, Istanbul Medeniyet University, Istanbul, Turkey. [2]Division of Systems Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College London, London, UK. [3]Department of Dermatology and Venereology, Istanbul Teaching and Research Hospital, Health Sciences University, Istanbul, Turkey. [4]Department of Dermatology and Venereology, Ege University, Istanbul, Turkey. ✉email: ozan.erdem@medeniyet.edu.tr; a.yilmaz23@imperial.ac.uk

assessment settings[19–24]. A smaller, but growing, body of work has also directly compared AI performance with that of medical students and residents across different specialties[25–30]. However, most benchmarks rely predominantly on text-only questions, leaving the capacity of these models to perform complex visual analysis and to recognize subtle clinical nuances insufficiently examined.[31] This limitation is particularly evident in visually intensive specialties such as dermatology, where diagnostic reasoning depends heavily on the interpretation and integration of image-based information[31–35].

Against this background, we conducted a faculty-curated evaluation of contemporary VLMs in undergraduate dermatology education using standardized clerkship examinations at a Turkish medical school. We compared four state-of-the-art models, GPT-4o and GPT-5 (both accessed via ChatGPT), Gemini 2.5 Flash, and Gemini 3 Pro, with fifth-year medical students on examinations combining text-only multiple-choice, multiple-select, and matching questions (60% of the total score) with image-based, structured open-ended questions (40%) spanning seven domains of dermatologic reasoning and practice. Using expert-validated answer keys, predefined scoring rubrics, and comparative statistical analyses, we aimed to characterize differences in text-only and image-based performance between VLMs and students and to explore domain-specific patterns within a real educational workflow, thereby providing insight into the current strengths and limitations of multimodal AI in dermatology.

Results

Exam difficulty levels and data distribution

To characterize baseline examination difficulty, student performance across the 10 dermatology clerkship exam sessions was analyzed. Mean student scores in the text-only component ranged from 78.46 ± 9.04 (April 2024) to 91.93 ± 4.66 (February 2024). In the image-based component, mean scores ranged from 72.19 ± 13.70 (December 2024) to 89.41 ± 6.43 (November 2023).

Differences in student performance between exam sessions were statistically significant for both text-only and image-based components (Kruskal–Wallis test, $p < 0.001$ for both), indicating variability in examination difficulty across sessions. Because standardized item-level difficulty classifications were not available, a complete item-level subgroup analysis could not be performed. Therefore, analyses were pooled across all 10 sessions, and examination-level difficulty was additionally assessed by correlating VLM performance with mean student scores for each session.

Performance on text-only examination

In the pooled text-only examination, all evaluated VLMs achieved higher mean scores than the medical student cohort (Table 1). Students achieved a mean score of 84.91 ± 9.04, whereas all VLMs exceeded 95 points on average (Fig. 1a).

A Kruskal–Wallis test demonstrated a significant difference among groups ($p < 0.001$). Post-hoc pairwise comparisons using the Mann–Whitney U test with Bonferroni correction showed that all VLMs (Gemini 3 Pro, Gemini 2.5 Flash, GPT-5, and GPT-4o) scored significantly higher than medical students ($p < 0.001$ for all comparisons).

Among the models, Gemini 3 Pro achieved the highest mean score (98.45 ± 1.42), which was significantly higher than GPT-4o (95.43 ± 1.98, $p < 0.001$) and GPT-5 (95.19 ± 2.27, $p < 0.001$). No statistically significant difference was observed between GPT-4o and GPT-5 in the text-only component ($p = 1.00$).

Performance on image-based examination

In the image-based examination, performance differed significantly among groups (Kruskal–Wallis test, $p < 0.001$; Table 1). Gemini 3 Pro achieved the highest mean score (93.94 ± 3.41), followed by GPT-5 (87.24 ± 6.97). Both models scored higher than the medical student cohort (82.94 ± 9.69), with statistically significant differences ($p < 0.001$ and $p = 0.015$, respectively).

Medical students achieved higher mean image-based scores than GPT-4o (77.36 ± 5.40) and Gemini 2.5 Flash (70.34 ± 5.08). These differences were statistically significant ($p < 0.001$ for both comparisons). Gemini 2.5 Flash had the lowest mean score among all groups in the image-based component (Fig. 1b).

Group	Text-only Score	Image-based Score	Overall Score
Gemini 3 Pro	98.45 ± 1.42	93.94 ± 3.41	96.64 ± 1.54
GPT-5	95.19 ± 2.27	87.24 ± 6.97	92.01 ± 2.69
GPT-4o	95.43 ± 1.98	77.36 ± 5.40	88.20 ± 2.01
Gemini 2.5 Flash	96.45 ± 1.42	70.34 ± 5.08	86.00 ± 2.17
Medical Students	84.91 ± 9.04	82.94 ± 9.69	84.12 ± 7.99

Table 1. Performance of vision–language models and medical students across text-only, image-based, and overall examination modalities. For VLMs, data are presented as mean scores ± standard deviation across run-level observations from ten examination sessions, with five repeated runs per session. For medical students, data are presented as mean scores ± standard deviation across individual student scores. Overall scores represent a weighted average of text-only (60%) and image-based (40%) components.



Fig. 1. Distribution of performance scores across text-only, image-based, and overall assessment modalities. Violin plots display score distributions for each participant group. The width of each violin represents kernel density estimation. Internal box plots indicate the median (central line), interquartile range (box limits), and whiskers (1.5× interquartile range). White diamond markers denote mean scores. **(a)** Text-only examination scores, **(b)** image-based examination scores, **(c)** overall examination scores combining text-only and image-based components.

Group	Multiple-choice	Multiple-selection	Matching
Gemini 3 Pro	98.52 ± 1.98	98.31 ± 2.35	98.40 ± 3.18
GPT-5	95.94 ± 3.04	96.41 ± 3.84	88.67 ± 6.77
GPT-4o	96.77 ± 2.52	94.92 ± 4.29	89.60 ± 5.82
Gemini 2.5 Flash	96.26 ± 2.29	97.23 ± 2.63	95.73 ± 4.90
Students	81.99 ± 10.73	89.45 ± 8.54	90.19 ± 10.35

Table 2. Accuracy of vision–language models and medical students across different text-only question formats. Data are presented as mean accuracy percentages ± standard deviation.

Analysis by text-only question type

Text-only performance was further analyzed according to question format (Table 2; Fig. 2a). In multiple-choice questions, Gemini 3 Pro achieved the highest accuracy (98.52%), followed by GPT-4o (96.77%), Gemini 2.5 Flash (96.26%), and GPT-5 (95.94%). Medical students achieved a mean accuracy of 81.99%. Differences between students and all VLMs were statistically significant ($p < 0.001$).

In multiple-selection questions, Gemini 3 Pro achieved the highest accuracy (98.31%), followed by Gemini 2.5 Flash (97.23%), GPT-5 (96.41%), and GPT-4o (94.92%). Medical students achieved a mean accuracy of 89.45%. All VLMs scored significantly higher than students ($p < 0.001$).

In matching questions, Gemini 3 Pro achieved the highest accuracy (98.40%), followed by Gemini 2.5 Flash (95.73%). Medical students achieved a mean accuracy of 90.19%, which was higher than GPT-4o (89.60%) and GPT-5 (88.67%). Differences between students and both GPT models were statistically significant ($p < 0.05$).

Analysis by image-based sub-domains

We further examined VLM performance across seven image-based dermatological sub-domains (Table 3; Fig. 2b). Because item-level rubric data were not available for medical students, image-based sub-domain analyses were restricted to VLMs; therefore, no direct student–model comparisons were performed at the sub-domain level.

Performance varied across sub-domains and models. Gemini 3 Pro achieved the highest or near-highest accuracy across all image-based sub-domains, particularly in descriptive reporting (99.40%), diagnosis (96.67%), etiology (95.00%), and general knowledge (97.92%). GPT-5 also performed strongly, with comparable results in descriptive reporting, diagnosis, and etiology, but lower accuracy in differential diagnosis and general knowledge. GPT-4o showed intermediate performance overall, whereas Gemini 2.5 Flash had the lowest accuracy in several reasoning-oriented sub-domains, including etiology, differential diagnosis, and treatment.



Fig. 2. Comparative performance of vision–language models and medical students across text-only and image-based assessment domains. **(a)** Mean accuracy percentages for text-only examinations stratified by question type (multiple choice, multiple selection, and matching). **(b)** Radar chart showing mean accuracy across seven image-based sub-domains: elementary lesion recognition, descriptive reporting, diagnosis, etiology, differential diagnosis, treatment, and general dermatological knowledge. Sub-domain–level image-based performance is shown for vision–language models only, as item-level rubric data were not available for medical students.

Sub-domain	Gemini 3 Pro	GPT-5	GPT-4o	Gemini 2.5 Flash
Elementary lesion recognition	89.40 ± 7.47	79.90 ± 12.92	79.60 ± 16.44	75.00 ± 15.75
Descriptive reporting	99.40 ± 3.14	99.00 ± 3.64	96.60 ± 8.48	77.60 ± 18.58
Diagnosis	96.67 ± 10.1	96.67 ± 8.65	90.00 ± 14.52	86.40 ± 16.38
Etiology	95.00 ± 15.1	94.80 ± 15.15	61.80 ± 28.76	53.00 ± 35.93
Differential diagnosis	87.80 ± 13.2	73.40 ± 23.33	62.67 ± 26.98	54.73 ± 24.93
Treatment	88.60 ± 7.56	86.20 ± 12.60	79.60 ± 10.68	58.60 ± 23.04
General knowledge	97.92 ± 4.14	86.24 ± 16.92	72.56 ± 11.14	72.48 ± 13.89

Table 3. Accuracy of vision–language models across distinct image-based dermatological sub-domains. Data are presented as mean accuracy percentages ± standard deviation.

Comparison of text-only and image-based performance

Within-group comparisons between text-only and image-based scores were performed using the Wilcoxon signed-rank test (Fig. 3). All groups demonstrated statistically significant differences between modalities ($p < 0.05$).

The largest mean score differences between text-only and image-based components were observed for Gemini 2.5 Flash (26.1 points, $p < 0.001$) and GPT-4o (18.0 points, $p < 0.001$). GPT-5 demonstrated a smaller difference of 8.0 points ($p < 0.001$), while Gemini 3 Pro showed the smallest difference among models (4.5 points, $p < 0.001$). Notably, medical students demonstrated the most balanced profile with a mean difference of 1.97 points between modalities ($p = 0.002$).

Correlation with examination difficulty

To assess the relationship between examination difficulty and group performance, mean scores of each VLM were correlated with mean student scores across the 10 exam sessions.

In the text-only component, no significant correlation was observed between student scores and VLM scores (Fig. 4a). In contrast, in the image-based component, positive correlations were observed between student performance and the performance of GPT-4o and Gemini 2.5 Flash, whereas Gemini 3 Pro maintained consistently high performance across sessions (Fig. 4b).

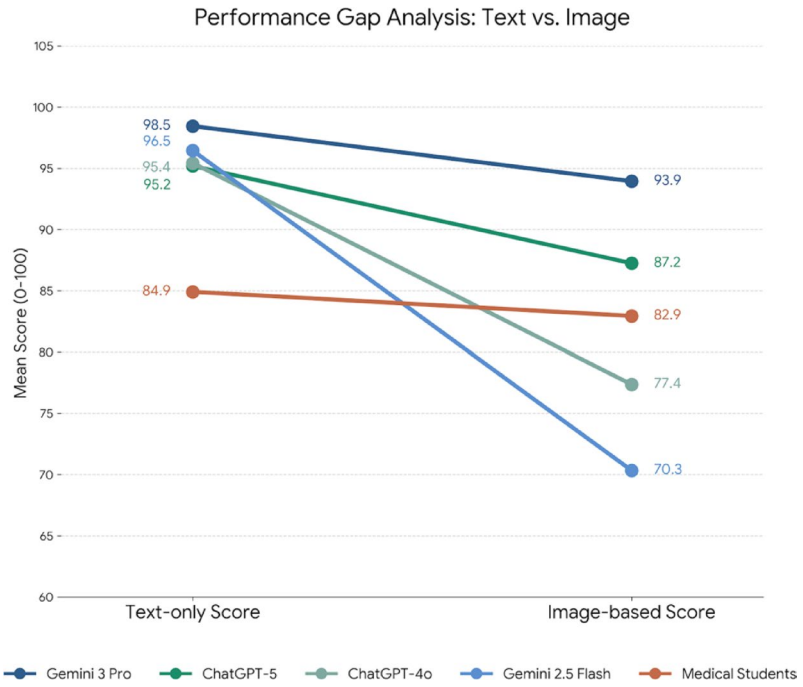


Fig. 3. Comparison of text-only and image-based performance within each participant group. Lines connect mean text-only scores (left) to mean image-based scores (right). The vertical distance between points represents the magnitude of score difference between modalities for each group.

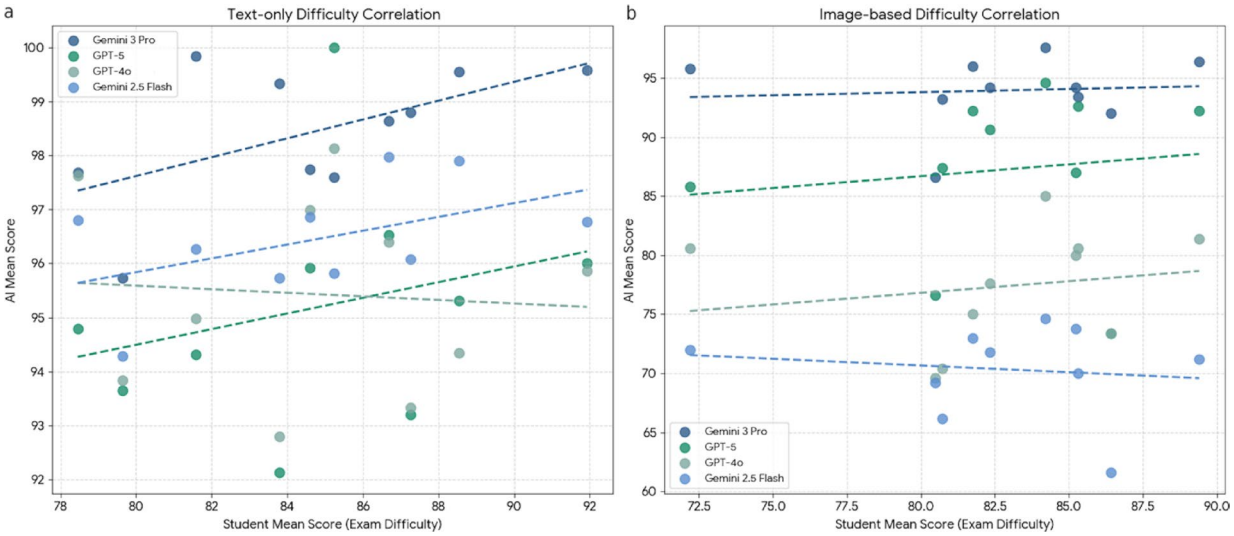


Fig. 4. Correlation between examination difficulty and group performance across 10 exam sessions. Scatter plots display the relationship between mean medical student scores for each exam session (x-axis) and corresponding mean scores of vision–language models (y-axis). **(a)** Text-only examination scores. **(b)** Image-based examination scores. Trend lines represent linear regression for each model.

Model output variability

Model output variability was assessed by analyzing score distributions across five repeated runs for each exam session (Fig. 5). Gemini 3 Pro demonstrated the lowest mean standard deviation across repeated runs in both text-only (SD = 0.72) and image-based (SD = 1.70) components, while higher variability was observed for Gemini 2.5 Flash, particularly in the image-based component (SD = 3.23). GPT-4o and GPT-5 demonstrated intermediate levels of variability.

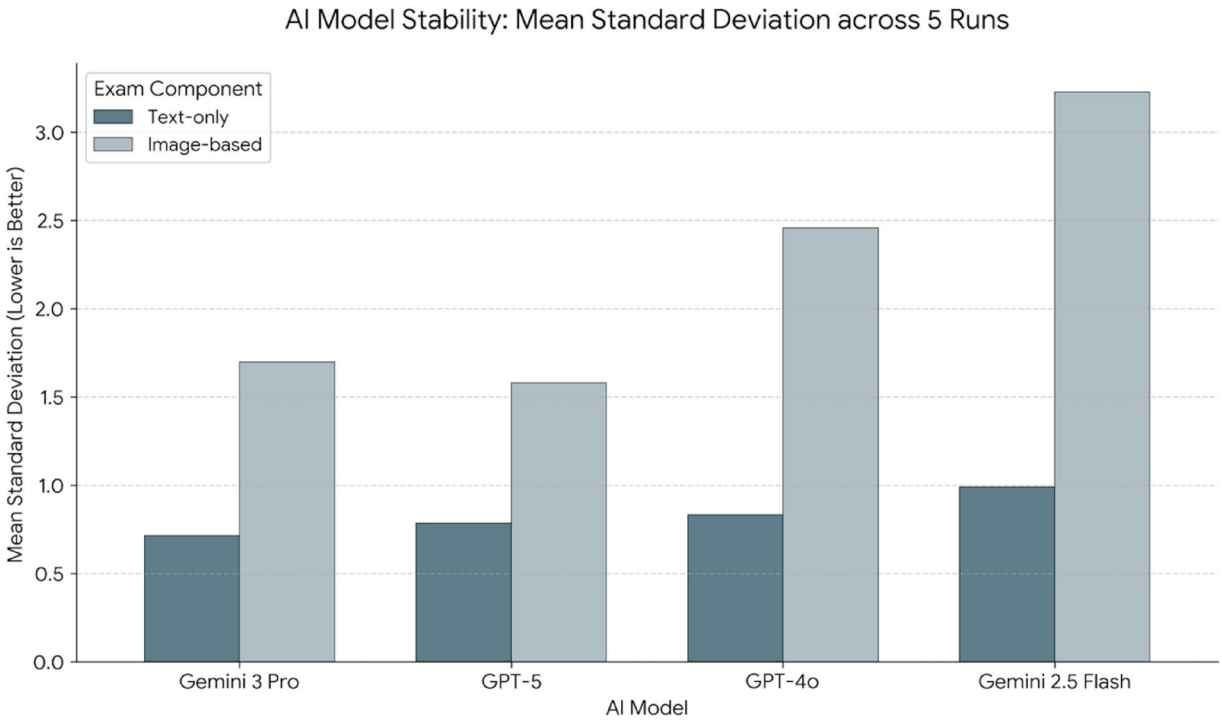


Fig. 5. Model output variability across repeated evaluations. Bar chart displays the mean standard deviation of scores across five repeated runs for each vision–language model in text-only and image-based examinations. Lower standard deviation values indicate lower score variability across runs.

Component	Gemini 3 Pro	GPT-5	GPT-4o	Gemini 2.5 Flash	Students
Text-only	98.36 ± 1.43	95.40 ± 2.21	94.79 ± 2.17	96.57 ± 1.00	84.91 ± 9.04
Image-based	94.20 ± 3.94	88.00 ± 6.67	78.90 ± 4.68	70.80 ± 4.24	82.94 ± 9.69
Overall	96.70 ± 2.04	92.44 ± 2.79	88.43 ± 1.99	86.26 ± 1.99	84.12 ± 7.99

Table 4. First-run sensitivity analysis comparing VLM and student performance across assessment components. Data are presented as mean scores ± standard deviation. For VLMs, only the first run from each examination session was included in this sensitivity analysis; repeated runs 2–5 were excluded.

First-run sensitivity analysis

Because most real-world interactions with VLMs involve receiving a single response to a given prompt, we performed a first-run sensitivity analysis to evaluate whether the main findings were driven by the inclusion of repeated model outputs in the primary pooled analysis (Table 4). In this analysis, only the first run of each VLM for each examination session was included, while the original single-attempt student scores were retained. The overall pattern of findings was broadly preserved. In the text-only component, all VLMs achieved higher mean scores than students. In the image-based component, Gemini 3 Pro and GPT-5 remained higher than the student cohort, whereas GPT-4o and Gemini 2.5 Flash scored below students. Overall scores also remained higher for all VLMs than for students.

Discussion

This study provides a structured comparison of contemporary VLMs and medical students within an authentic undergraduate dermatology assessment framework, revealing clear modality-dependent differences in performance. By analyzing both text-only and image-based examination components across multiple exam sessions, we found that strong performance in text-based dermatologic assessments does not consistently translate to image-based reasoning. While several VLMs achieved near-ceiling scores on text-only tasks, performance diverged substantially on image-based questions, with marked variability across model architectures. These findings indicate that multimodal competence in dermatology remains heterogeneous and task-dependent, underscoring the importance of evaluating visual and textual reasoning separately rather than inferring overall capability from text-based performance alone.

In the text-only domain, all evaluated VLMs achieved near-ceiling performance, exceeding the mean scores of the medical student cohort across exam sessions. Moreover, model performance remained largely invariant to fluctuations in exam difficulty that significantly affected student scores. These findings are consistent with prior

studies demonstrating that LLMs can reach or exceed passing thresholds on national licensing and specialty board examinations, reflecting strong capabilities in factual recall and structured medical knowledge application[14,36,37]. However, aggregate text-based scores obscure meaningful differences between question formats. While VLMs consistently outperformed students on multiple-choice and multiple-select items, students achieved comparable or slightly higher accuracy than GPT-4o and GPT-5 on matching-type questions, suggesting that certain associative reasoning tasks without contextual scaffolding may still favor human cognitive strategies.

The image-based examination component revealed a markedly different performance hierarchy. While Gemini 3 Pro and GPT-5 achieved higher mean scores than medical students, students significantly outperformed GPT-4o and Gemini 2.5 Flash. This reversal underscores that visual diagnostic competence remains a challenging domain for some VLMs, despite their strong performance on text-based tasks. Similar modality-dependent performance patterns have been reported in recent evaluations, where accuracy declined for items incorporating images or complex visual contexts[23,38,39]. Notably, the student cohort exhibited the smallest performance difference between text-only and image-based tasks, indicating a relatively consistent integration of theoretical knowledge and visual interpretation. In contrast, earlier or lighter model variants showed substantial performance declines when transitioning from text to image-based tasks, reflecting limited multimodal consistency. Consequently, text-based accuracy alone is an incomplete proxy for multimodal competence and should be interpreted cautiously when evaluating the readiness of VLMs for visually intensive clinical domains such as dermatology[40].

Sub-domain analyses further clarify the nature of this discrepancy. While most models performed well in tasks emphasizing visual description and pattern recognition, notable limitations emerged in domains requiring etiological reasoning from visual input. For example, GPT-4o and Gemini 2.5 Flash demonstrated reduced accuracy in etiology-related questions despite achieving moderate diagnostic accuracy. This pattern suggests that correct visual labeling does not necessarily coincide with coherent pathophysiological reasoning, a distinction that has also been emphasized in prior analyses of AI-assisted dermatologic diagnosis[41]. In contrast, Gemini 3 Pro and GPT-5 maintained higher performance across etiological and treatment-related domains, indicating improved integration of visual features with underlying clinical knowledge.

Model reliability represents an additional educationally and clinically relevant dimension. Repeated exam administrations revealed substantial variability in output stability across models. Gemini 3 Pro exhibited the lowest performance variance across both modalities, whereas Gemini 2.5 Flash showed greater stochastic variability, particularly in image-based tasks. Importantly, such variability may limit the interpretability and pedagogical reliability of model outputs in educational settings[42]. To determine whether the main findings were driven by the inclusion of repeated model outputs in the primary pooled analysis, we additionally performed a first-run sensitivity analysis, which more closely reflects a single-response use scenario. The overall performance pattern was broadly preserved, suggesting that the main results were not solely attributable to the inclusion of repeated model outputs.

Several methodological strengths enhance the relevance of this study, including the use of original non-English examination materials, a large human cohort, and repeated AI evaluations that enabled aggregate benchmarking and assessment of performance variability. However, important limitations should be acknowledged. This was a single-center study based on one institutional curriculum, and the examinations, although clinically oriented, remained structured academic assessments. Moreover, students completed each examination once under real examination conditions, whereas VLMs were evaluated across repeated runs; therefore, pooled aggregate model performance is not directly equivalent to single-attempt human performance and may partially smooth inconsistent responses or occasional errors. Static images also cannot capture the full complexity of dermatologic evaluation, which may involve dermoscopy, palpation, and longitudinal clinical context. In addition, because image-based open-ended VLM responses were not scored through blinded independent parallel assessment, inter-rater reliability statistics could not be calculated, and observer bias cannot be excluded. Finally, student item-level rubric scores for the image-based component were not retained in the retrospective records, preventing sub-domain analysis of student performance. Accordingly, image-based sub-domain findings should be interpreted as descriptive within-model and between-model patterns rather than direct student–model comparisons.

In conclusion, selected VLMs achieved high aggregate accuracy on text-only dermatology questions, whereas image-based performance was more heterogeneous and strongly model-dependent. Under the repeated-run evaluation protocol, GPT-5 and Gemini 3 Pro achieved higher aggregate scores than the student cohort mean in the image-based component, while GPT-4o and Gemini 2.5 Flash scored lower. Medical students showed a smaller gap between text-only and image-based performance, suggesting relatively greater cross-modal consistency within this structured examination setting. However, these findings should be interpreted cautiously, given the differences between single-attempt student testing and repeated model evaluations, as well as the observed variability in model outputs. They should not be considered evidence of real-world clinical diagnostic competence or direct superiority over medical students. Rather, the results highlight the need for modality-aware and reliability-sensitive benchmarking and suggest that VLMs may have potential as complementary, human-supervised tools in dermatology education rather than as standalone systems.

Methods

Study design and setting

This comparative cross-sectional study assessed the performance of VLMs and fifth-year medical students on standardized dermatology clerkship examinations at Istanbul Medeniyet University Faculty of Medicine. The dataset comprised ten consecutive clerkship rotations conducted between September 2023 and January 2025. Both the text-only and image-based examinations were mandatory components of the official clerkship

assessment. All analyses were performed using the original Turkish questions and responses to evaluate performance in the language of instruction.

Ethics approval and clinical image source

The study was approved by the Prof. Dr. Süleyman Yalçın City Hospital Non-Interventional Clinical Research Ethics Committee 6 November 2025, with approval number 2025/0263. The study was conducted in accordance with the principles of the Declaration of Helsinki.

All clinical images used in the image-based examination were obtained from the anonymized clinical photography archive of the Department of Dermatology and Venereology, Istanbul Medeniyet University, and had been incorporated into routine clerkship examination materials before the present study. No third-party or commercially licensed images were used. The use of these images for this study and manuscript was approved by the Ethics Committee and the Dean's Office of the Faculty of Medicine.

Participants

A total of 252 fifth-year medical students participated across ten clerkship groups (21–29 students per group; mean ≈ 25). Because the examinations were a required part of the rotation, no exclusion criteria were applied.

Exam structure and scoring

Across the ten exam sessions, the dataset included 500 text-only items and 200 image-based items. Each session consisted of:

- Text-only examination (50 items): 31 single-answer multiple-choice (Q1–31), 13 multiple-select (Q32–44), and 6 matching questions (Q45–50).
- Image-based examination (20 items): structured open-ended questions spanning seven sub-domains: elementary lesion recognition, descriptive reporting, diagnosis, etiology, differential diagnosis, treatment, and general dermatological knowledge.

Text-only items were graded using predefined answer keys, whereas image-based items were graded using weighted rubrics. The overall examination score was calculated as a weighted combination of modalities (text-only 60%, image-based 40%). The detailed exam structure and scoring framework is summarized in Table 5.

AI models and testing procedure

Four VLMs were evaluated using their standard web-based user interfaces: GPT-4o and GPT-5 (accessed via ChatGPT) from OpenAI, and Gemini 2.5 Flash and Gemini 3 Pro from Google. These systems were selected as widely accessible, general-purpose commercial VLMs available through user-facing web interfaces. Model selection also reflected availability during the study period: GPT-4o and Gemini 2.5 Flash were evaluated during the initial analysis phase in May 2025, whereas GPT-5 and Gemini 3 Pro were evaluated in October 2025 after newer models from both providers became available. Because these web interfaces did not allow user-level control or recording of temperature and related sampling parameters, the models were evaluated using the default generation settings available at the time of testing.

Exam component	Subcategory (question range)	No. of items	Scoring scheme	Max score
Text-only exam	Multiple-choice (Q1–31)	31	2 points per correct answer	62
	Multiple-select (Q32–44)	13	(Correct selections ÷ 3) × 2; max 2 points per item; invalid (0 points) if > 3 choices selected	26
	Matching (Q45–50)	6	(Correct matches ÷ 5) × 2; max 2 points per item	12
	Total	50	—	100
Image-based exam	Elementary lesion recognition (Q1–4)	4	Weighted rubric; max 5 points per item	20
	Descriptive reporting (Q5–6)	2	Weighted rubric; max 5 points per item	10
	Diagnosis (Q7–9)	3	Single-answer; max 5 points per item	15
	Etiology (Q10–11)	2	Single-answer; max 5 points per item	10
	Differential diagnosis (Q12–13)	2	Weighted rubric; max 5 points per item	10
	Treatment (Q14–15)	2	Weighted rubric or single-answer; max 5 points per item	10
	General dermatological knowledge (Q16–20)	5	Weighted rubric or single-answer; max 5 points per item	25
	Total	20	—	100

Table 5. Structure and scoring scheme of the dermatology clerkship examinations. Text-only and image-based components were each scaled to a maximum of 100 points. For multiple-select items, partial credit was calculated as (number of correct selections ÷ 3) × 2 (maximum 2 points per item); responses selecting more than three options were scored as 0. For matching items, partial credit was calculated as (number of correct matches ÷ 5) × 2 (maximum 2 points per item). 'Weighted rubric' indicates predefined point allocations (up to 5 points per item) based on faculty-approved grading criteria (e.g., awarding credit for specific expected elements within an open-ended response). Overall score was computed as a weighted average: 0.60 × text-only score + 0.40 × image-based score.

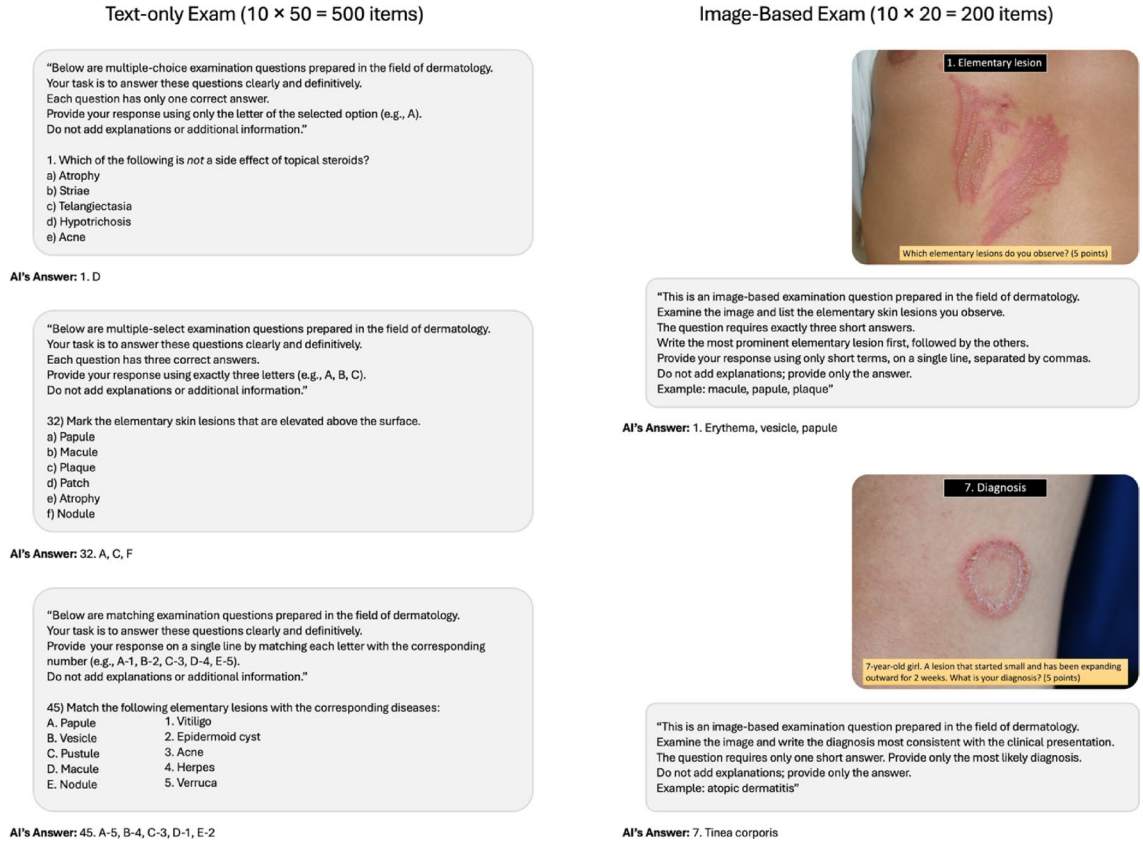


Fig. 6. Representative examples from the text-only (left) and image-based (right) components of the dermatology clerkship examinations, illustrating the standardized prompt templates and example model outputs. Prompt templates were format-specific and instructed models to return answers in a constrained output format (e.g., letter-only responses for text-only items and short, structured responses for image-based items). All analyses were performed using the original Turkish questions and responses; English translations are provided for illustrative purposes only.

Model testing used the standardized, format-specific Turkish prompts provided in Supplementary Information I. The same prompts were applied verbatim across all models and examination items. To prevent contamination from previous interactions, each prompt was submitted in a new temporary chat session without access to prior conversation history. Text-only items were submitted in grouped batches according to question format, with explicit response-format instructions for multiple-choice, multiple-select, and matching items. Image-based items were submitted individually and required structured open-ended responses with predefined constraints on the type and number of expected answers. Clinical images were presented to the models exactly as they appeared in the examination files administered to students, without additional preprocessing, cropping, enhancement, or modification.

All model outputs were manually recorded at the item level in the same examination score spreadsheet structure used for student results. Each model response was stored separately, and item-level scores were assigned according to the predefined answer keys and grading rubrics. To account for stochastic variability in generative outputs, each examination session was repeated five times per model using identical prompts and input materials. Run-level model scores were retained and used as repeated model observations in the pooled performance analyses. Score distributions across repeated runs were additionally analyzed to assess output variability.

For transparency, standardized prompt templates and representative examples are provided in Supplementary Information I and illustrated in Fig. 6. The full content of the first evaluated text-only examination (22 September 2023), including English translations and the corresponding expert-validated answer key, is provided in Supplementary Information II. The image-based examination file containing clinical photographs is not included as a publicly available supplementary file because explicit study-specific consent for public publication was not obtained from each individual patient. Access to these image-based examination materials may be made available from the corresponding author upon reasonable request, subject to institutional and ethical restrictions.

Gold-standard answers and evaluation of model outputs

Faculty members from relevant subspecialty areas authored the examination items before the present study as part of routine undergraduate dermatology clerkship assessment. All questions were compiled, reviewed, and approved by the clerkship coordinator. Gold-standard answers for text-only items and grading rubrics for

image-based items were prepared by the item authors and verified by the clerkship coordinator. No chatbot, LLM, VLM, or other generative AI system was used to generate, revise, select, or validate the examination questions, clinical images, answer keys, or scoring rubrics.

VLM responses were evaluated against the same predefined answer keys and grading rubrics used for student assessment. Text-only responses were scored according to the corresponding answer keys, whereas structured open-ended image-based responses were scored item by item according to the faculty-approved weighted rubrics. VLM responses were scored by the clerkship coordinator and a senior dermatology resident through collaborative review with task sharing, reflecting the routine grading workflow used in the clerkship. No independent parallel scoring, averaging, or formal consensus procedure was applied.

Availability of student image-based sub-domain data

Student image-based data were obtained retrospectively from official clerkship records. While total image-based examination scores were available for all students, sub-domain–level rubric components were not recorded in the examination system at the time of assessment. Accordingly, sub-domain analyses for image-based performance could be conducted only for VLMs, for which rubric-based scoring was prospectively recorded at the item level.

Statistical analysis

All statistical analyses were performed in Python (version 3.10) using the *pandas*, *numpy*, and *scipy* libraries. Data visualization was conducted using *matplotlib* and *seaborn* to generate violin plots, radar charts, and slope graphs. Normality of score distributions was assessed using the Shapiro–Wilk test. As several variables deviated from normality, non-parametric statistical methods were applied. Overall between-group differences were evaluated using the Kruskal–Wallis test, followed by pairwise Mann–Whitney U tests for relevant group comparisons (students versus each model and between-model comparisons), with Bonferroni correction applied for multiple testing. To assess within-group differences between text-only and image-based performance (the modality gap), the Wilcoxon signed-rank test was used. Correlations between mean student scores and model scores across examination sessions were examined using Spearman's rank correlation coefficient. Model output variability was quantified descriptively using the standard deviation of total scores across five repeated runs per model and examination session. Statistical significance was defined as a two-sided p value < 0.05.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 30 December 2025; Accepted: 25 August 2026
Published online: 05 September 2026

References

1. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
2. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
3. Gordon, M. et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. *Med. Teach.* **46**, 446–470 (2024).
4. Vrdoljak, J., Boban, Z., Vilović, M., Kumrić, M. & Božić, J. A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. *Healthcare* **13**, 603 (2025).
5. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
6. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
7. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
8. Hartsock, I. & Rasool, G. Vision-language models for medical report generation and visual question answering: a review. *Front. Artif. Intell.* **7**, 1430984 (2024).
9. Krones, F., Marikkar, U., Parsons, G., Szmul, A. & Mahdi, A. Review of multimodal machine learning approaches in healthcare. *Inf. Fusion*. **114**, 102690 (2025).
10. Wang, Z. et al. A perspective for adapting generalist AI to specialized medical AI applications and their challenges. *Npj Digit. Med.* **8**, 429 (2025).
11. Yan, S. et al. A multimodal vision foundation model for clinical dermatology. *Nat. Med.* **31**, 2691–2702 (2025).
12. Zhou, J. et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nat. Commun.* **15**, 5649 (2024).
13. Yilmaz, A. et al. Resource-efficient medical vision language model for dermatology via a synthetic data generation framework. 05.17.25327785 Preprint at (2025). https://doi.org/10.1101/2025.05.17.25327785 (2025).
14. Kung, T. H. et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit. Health*. **2**, e0000198 (2023).
15. Gilson, A. et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med. Educ.* **9**, e45312 (2023).
16. Brin, D. et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci. Rep.* **13**, 16492 (2023).
17. Sadeq, M. A. et al. AI chatbots show promise but limitations on UK medical exam questions: a comparative performance study. *Sci. Rep.* **14**, 18859 (2024).
18. Casals-Farre, O. et al. Assessing ChatGPT 4.0's Capabilities in the United Kingdom Medical Licensing Examination (UKMLA): A Robust Categorical Analysis. *Sci. Rep.* **15**, 13031 (2025).
19. Kim, H. J. et al. Performance evaluation of large language models on Korean medical licensing examination: a three-year comparative analysis. *Sci. Rep.* **15**, 36082 (2025).
20. Jung, L. B. et al. ChatGPT Passes German State Examination in Medicine With Picture Questions Omitted. *Dtsch. Arzteblatt Int.* **120**, 373–374 (2023).
21. Takagi, S., Watari, T., Erabi, A. & Sakaguchi, K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. *JMIR Med. Educ.* **9**, e48002 (2023).
22. Rojas, M., Rojas, M., Burgess, V., Toro-Pérez, J. & Salehi, S. Exploring the Performance of ChatGPT Versions 3.5, 4, and 4 With Vision in the Chilean Medical Licensing Examination: Observational Study. *JMIR Med. Educ.* **10**, e55048–e55048 (2024).

23. Miyazaki, Y. et al. Performance of ChatGPT-4o on the Japanese Medical Licensing Examination: Evalution of Accuracy in Text-Only and Image-Based Questions. *JMIR Med. Educ.* **10**, e63129–e63129 (2024).
24. Luo, D. et al. Evaluating the performance of GPT-3.5, GPT-4, and GPT-4o in the Chinese National Medical Licensing Examination. *Sci. Rep.* **15**, 14119 (2025).
25. Guerra, G. A. et al. GPT-4 Artificial Intelligence Model Outperforms ChatGPT, Medical Students, and Neurosurgery Residents on Neurosurgery Written Board-Like Questions. *World Neurosurg.* **179**, e160–e165 (2023).
26. Massey, P. A., Montgomery, C. & Zhang, A. S. Comparison of ChatGPT–3.5, ChatGPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations. *J. Am. Acad. Orthop. Surg.* **31**, 1173–1179 (2023).
27. Roos, J., Kasapovic, A., Jansen, T. & Kaczmarczyk, R. Artificial Intelligence in Medical Education: Comparative Analysis of ChatGPT, Bing, and Medical Students in Germany. *JMIR Med. Educ.* **9**, e46482 (2023).
28. Meyer, A., Riese, J. & Streichert, T. Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the Written German Medical Licensing Examination: Observational Study. *JMIR Med. Educ.* **10**, e50965 (2024).
29. Bahir, D. et al. Gemini AI vs. ChatGPT: A comprehensive examination alongside ophthalmology residents in medical knowledge. *Graefes Arch. Clin. Exp. Ophthalmol.* https://doi.org/10.1007/s00417-024-06625-4 (2024).
30. Zengin, A., Ulfanov, O., Bag, Y. M. & Ulas, M. Artificial Intelligence Versus Medical Students in General Surgery Exam. *Indian J. Surg.* **87**, 68–73 (2025).
31. Behrmann, J. et al. Chat generative pre-trained transformer's performance on dermatology-specific questions and its implications in medical education. *J. Med. Artif. Intell.* **6**, 16–16 (2023).
32. Passby, L., Jenko, N. & Wernham, A. Performance of ChatGPT on Specialty Certificate Examination in Dermatology multiple-choice questions. *Clin. Exp. Dermatol.* **49**, 722–727 (2024).
33. Fan, K. S. & Fan, K. H. Dermatological Knowledge and Image Analysis Performance of Large Language Models Based on Specialty Certificate Examination in Dermatology. *Dermato* **4**, 124–135 (2024).
34. Göçer Gürok, N. & Öztürk, S. The Performance of AI in Dermatology Exams: The Exam Success and Limits of ChatGPT. *J. Cosmet. Dermatol.* **24**, e70244 (2025).
35. Atılan, A. U. & Çetin, N. Benchmarking Large Language Models on the Turkish Dermatology Board Exam: A Comparative Multilingual Analysis. *Turk. J. Dermatol.* **19**, 126–133 (2025).
36. Bicknell, B. T. et al. ChatGPT-4 Omni Performance in USMLE Disciplines and Clinical Skills: Comparative Analysis. *JMIR Med. Educ.* **10**, e63430 (2024).
37. Brin, D. et al. How GPT models perform on the United States medical licensing examination: a systematic review. *Discov Appl. Sci.* **6**, 500 (2024).
38. Yang, X. & Chen, W. The performance of ChatGPT on medical image-based assessments and implications for medical education. *BMC Med. Educ.* **25**, 1192 (2025).
39. Liu, M. et al. Evaluating the Effectiveness of advanced large language models in medical Knowledge: A Comparative study using Japanese national medical examination. *Int. J. Med. Inf.* **193**, 105673 (2025).
40. Jin, Q. et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *Npj Digit. Med.* **7**, 190 (2024).
41. Yu, Z., Xin, C., Yu, Y., Xia, J. & Han, L. AI dermatology: Reviewing the frontiers of skin cancer detection technologies. *Intell. Oncol.* **1**, 89–104 (2025).
42. Han,H. Challenges of reproducible AI in biomedical data science. *BMC Med. Genomics*. **18** (8) (2025). https://doi.org/10.1186/s12920-024-02072-6

Acknowledgements

Abdurrahim Yilmaz has been funded by the President's PhD Scholarships at Imperial College London.

Author contributions

O.E. and A.Y. conceptualized and designed the study and developed the methodology. Formal analysis was performed by O.E., A.S.Ş., and A.Y. Investigation was carried out by O.E. and A.S.Ş. Resources were provided by O.E., V.A.E., M.A.K., and M.S.G. O.E. prepared the visualizations. The original draft of the manuscript was written by O.E. and A.Y. Manuscript review and editing were performed by E.G. and B.B.D. M.S.G. supervised the study. All authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval

The study was approved by the Prof. Dr. Süleyman Yalçın City Hospital Non-Interventional Clinical Research Ethics Committee on 6 November 2025 (approval no. 2025/0263) and was conducted in accordance with the principles of the Declaration of Helsinki. The requirement for individual informed consent was waived by the Ethics Committee because the study involved retrospective analysis of de-identified examination data collected as part of routine educational assessment. Permission to use the student examination data for research purposes was obtained from the Dean's Office of the Faculty of Medicine.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-68943-3.

Correspondence and requests for materials should be addressed to O.E. or A.Y.

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

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026