Exploring the role of Large Language Models in Melanoma: a Systemic Review

Mor Zarfati 1, Girish N Nadkarni 2, Benjamin S Glicksberg 2, Moti Harats 3, Shoshana Greenberger 4, Eyal Klang 2*, Shelly Soffer 5*

1 Department of Internal Medicine, Soroka University Medical Center, Beer-Sheva, Israel.
2 Division of Data-Driven and Digital Medicine, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, United States.
3 Department of Plastic and Reconstructive Surgery, Sheba Medical Center, Ramat-Gan, affiliated with the School of Medicine, Tel-Aviv University, Tel-Aviv, 52621, Israel.
4 Department of Dermatology, Pediatric Dermatology Unit, Sheba Medical Center, Ramat Gan, Israel; Faculty of Medicine, Tel Aviv University, Tel Aviv, 52621, Israel.
5 Institute of Hematology, Davidoff Cancer Center, Rabin Medical Center; Petah-Tikva, Israel.
* equal contribution

ABSTRACT

Background: Large language models (LLMs) are gaining recognition across various medical fields; however, their specific role in dermatology, particularly in melanoma care, is not well-defined. This systematic review evaluates the current applications, advantages, and challenges associated with the use of LLMs in melanoma care.

Methods: We conducted a systematic search of PubMed and Scopus databases for studies published up to July 23, 2024, focusing on the application of LLMs in melanoma. Identified studies were categorized into three subgroups: patient education, diagnosis and clinical management. The review process adhered to PRISMA guidelines, and the risk of bias was assessed using the modified QUADAS-2 tool.

Results: Nine studies met the inclusion criteria. Five studies compared various LLM models, while four focused on ChatGPT. Three studies specifically examined multi-modal LLMs. In the realm of patient education, ChatGPT demonstrated high accuracy, though it often surpassed the recommended readability levels for patient comprehension. In diagnosis applications, multi-modal LLMs like GPT-4V showed capabilities in distinguishing melanoma from benign lesions. However, the diagnostic accuracy varied considerably, influenced by factors such as the quality and diversity of training data, image resolution, and the models' ability to integrate clinical context. Regarding management advice, one study found that ChatGPT provided more reliable

management advice compared to other LLMs, yet all models lacked depth and specificity for individualized decision-making.

Conclusions: LLMs, particularly multimodal models, show potential in improving melanoma care through patient education, diagnosis, and management advice. However, current LLM applications require further refinement and validation to confirm their clinical utility. Future studies should explore fine-tuning these models on large dermatological databases and incorporate expert knowledge.

INTRODUCTION

Large language models (LLMs), including ChatGPT, Gemini and Llama, are artificial intelligence (AI) models designed to understand and generate human-like text.[1] These models are gaining recognition across various medical specialties for their potential to assist with clinical tasks.[2–7] However, their specific role in dermatology, particularly in melanoma care, remains under investigation.[8] Multi-modal LLMs, such as GPT-4 Vision (GPT-4V), further expand this potential by combining visual and textual data. This capability could improve applications in medical imaging and diagnosis.[9]

Previous studies have shown mixed results, leading to caution among dermatologists regarding the use of these models.[10] Nevertheless, with appropriate optimization, LLMs may improve melanoma diagnosis, patient communication, and treatment outcomes.

This systematic review aims to evaluate the current applications, advantages, and challenges associated with the use of LLMs in melanoma care.

FOUNDATIONAL CONCEPTS

Below are the key concepts related to LLMs and their applications in healthcare. In **Figure 1**, we present a hierarchy diagram of AI terms.

Artificial Intelligence and Deep Learning

AI refers to the development of algorithms capable of performing tasks that typically require human intelligence. Examples include language comprehension and image pattern recognition. Deep learning is a subset of AI that employs artificial neural networks to analyse different types of data and learn from it.[11,12]

Artificial Neural Networks

Artificial neural networks form the foundation of deep learning. Inspired by biological neural networks, they consist of interconnected nodes, or "neurons," organized in layers. Each neuron receives inputs, processes them, and passes an output to the subsequent layer. Each neuron is a simple computational unit, similar to a single logistic regression function. By adjusting the connections between neurons based on the input data, neural networks can learn to recognize patterns and generate predictions.[11]

Large Language Models

LLMs are large deep learning models that process and generate human-like text. Composed of multiple transformer layers, these models employ an attention mechanism to selectively focus on different parts of the input data. This structure allows them to excel in tasks such as text recognition, language translation, and content generation.[13] Notable examples of LLMs include ChatGPT by OpenAI and LLaMA by Meta.

Multimodal Large Language Models

Multimodal LLMs extend the capabilities of traditional LLMs by incorporating multiple data modalities, such as text and images. These advanced models can analyse and interpret both

visual and textual information, making them particularly valuable in fields like radiology and dermatology, where accurate diagnosis often requires the synthesis of information from diverse sources.[14] In **Figure 2**, we present a diagram of possible uses of multimodal LLMs in dermatology.

METHODS

Search Strategy:

A systematic review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement (PRISMA) guidelines and the recommendations for systematic reviews of prediction models (CHARMS checklist). [15,16] The study is registered with PROSPERO (CRD42024575859).[17]

We searched the literature for applications of LLMs in melanoma using PubMed and Scopus. A systematic search of the published literature was conducted on July 23, 2024. Our search query was “(("Melanoma") AND (("ChatGPT") OR ("large language models") OR ("OpenAI") OR ("Microsoft Bing") OR ("google bard") OR ("google gemini")))”. To ensure thoroughness, we also reviewed the reference lists of relevant articles, but this did not yield any additional studies that met the inclusion criteria.

We excluded articles that did not specifically evaluate the application of LLMs in melanoma, non-original articles, and conference abstracts.

Study Selection:

The titles and abstracts of the identified studies were screened to determine their eligibility based on the inclusion and exclusion criteria. Any uncertainty was resolved through discussion between two reviewers, with a third reviewer consulted when necessary. The full texts of the

selected articles were then independently assessed by two reviewers (MZ, SS). Discrepancies were resolved through consensus or consultation with a third reviewer (EK).

Data Extraction:

Data extraction was conducted using a standardized form to ensure consistency. Key information extracted included the first author's name, year of publication, sample size, LLM model types, objectives, and main findings.

To investigate the specific applications and effectiveness of LLMs in different aspects of melanoma care, we divided the articles into three subgroups: patient education, clinical management, and diagnosis.

Quality Assessment and Risk of Bias

To evaluate the risk of bias, we used the adapted version of the Quality Assessment of Diagnostic Accuracy Studies criteria (QUADAS-2).[18]

RESULTS:

Our literature search yielded a total of 45 articles from PubMed and Scopus. After the removal of 9 duplicates, the screening process found 9 studies that met our inclusion criteria. We did not identify additional studies via reference screening.[19–27] The process of study selection and the screening methodology are detailed in the PRIZMA flow chart **(Figure 3).**

According to the QUADAS-2 tool, most papers scored as having a low to moderate risk of bias for the interpretation of the index test. A detailed assessment of the risk of bias is provided in **Supplementary Table 1.**

The characteristics of the studies are presented in **Table 1**. A summary of the objective, sample size, reference standard, main findings and conclusions are presented in **Table 2**. The main advantages and challenges in the included studies are presented in **Table 3**.

Of the nine studies, five were comparative, evaluating and comparing various LLM models, such as ChatGPT, BARD, and BingAI.[19,21,22,25,26] The remaining four studies focused on a single LLM, specifically different versions of ChatGPT.[20,23,24,27] Three studies specifically examined multimodal LLMs, such as GPT-4V and LLaVA, highlighting their unique capabilities and associated challenges.[21,23,25]

The included studies were diverse in their objectives, methodologies, and evaluation metrics. The studies focused on the application of LLMs in melanoma diagnosis, patient education, and clinical decision-making.

Patient education

Four studies evaluated the use of LLMs in patient education, focusing on the accuracy of responses to common patient questions.[20,22,26,27] ChatGPT 4.0 and ChatGPT 3.5 were noted for their relatively high accuracy.

Deliyannis et al. found that while both ChatGPT and BARD can generate accurate educational responses, both ChatGPT 4.0 and 3.5 outperformed BARD.[22] Anguita et al. focused on choroidal melanoma and found no significant accuracy differences between ChatGPT 3.5, Bing AI, and DocsGPT beta.[26]

Young et al. reported that ChatGPT 4.0 generates mostly accurate responses, scoring 4.9/5. However, only 64% of these responses were considered suitable for patient use, indicating that ChatGPT may be more effective as a supplemental tool in clinical practice. The study also found

that the average readability score corresponded to a college-level comprehension, suggesting that the content might be too advanced for public use.[27]

Roster et al. addressed this readability issue by evaluating ChatGPT's responses to questions about sunscreen and melanoma from the American Academy of Dermatology's (AAD) website. They investigated whether prompt engineering techniques (strategic prompting) could improve readability. The study compared ChatGPT's responses after two rounds of strategic prompting with the original answers from the AAD website. The findings showed that the initial prompt did not lower the reading level compared to the AAD content. However, with additional prompting, the reading level was reduced to 7th grade, compared to the AAD's 9th grade level. This suggests that with proper prompt engineering, LLMs could improve the readability of medical information for melanoma patients.[20]

Melanoma Diagnosis

Four studies examined the use of LLMs in melanoma diagnosis, focusing on their ability to identify and classify melanoma using clinical and dermoscopic data.[21,23–25] Multi-modal LLMs, such as GPT-4V and LLaVA, played a key role in the majority of these evaluations.

Cirone et al. assessed GPT-4V and LLaVA, emphasizing their ability to integrate visual and textual data. GPT-4V demonstrated superior performance, with an overall accuracy of 85%, compared to 45% for LLaVA. Notably, LLaVA had difficulty recognizing melanoma in skin of color, unlike GPT-4V.[25] This finding is consistent with those of Akrout et al., who also showed that GPT-4V outperformed LLaVA across all assessed features, though both models require further refinement to enhance diagnostic accuracy.[21]

This suggests that ChatGPT Vision may not yet be suitable for independent clinical use without additional refinement.

Management advice

Only one study specifically evaluated the use of LLMs in providing melanoma management advice. Mu et al. conducted a comparative analysis of several LLMs (ChatGPT 4.0, BARD and BingAI) to assess their performance in this context. The study used five prompts related to melanoma

management. ChatGPT 4.0 consistently provided more reliable, evidence-based clinical advice, outperforming the other models, with significant differences noted compared to BARD and marginally compared to BingAI. However, none of the models evaluated the risks and benefits associated with their recommendations. The limited number of questions restricts the generalizability of the findings. [19]

DISCUSSION:

This review's findings underscore the potential of LLMs across various domains in melanoma care, including patient education, disease diagnosis and management advice. Of particular interest is the emergence of multi-modal LLMs, which integrate visual and textual data to address the complexities of medical imaging and clinical decision-making.

In patient education, LLMs demonstrated ability to generate accurate and readable responses to common melanoma-related queries. For example, Roster et al. showed that strategic prompting can enhance the readability of ChatGPT's outputs.[20] This finding suggests that with appropriate fine-tuning, LLMs could become valuable tools for creating accessible patient education materials, enabling individuals to make informed decisions.

In melanoma diagnosis, multi-modal LLMs such as GPT-4V and LLaVA exhibited capabilities in distinguishing melanoma from benign lesions. Cirone et al. and Akrout et al. demonstrated GPT-4V's superior performance,[21,25] particularly in handling variations in skin tone and image manipulations.[25] Zhou et al. presented SkinGPT-4, a multi-modal LLM trained on a large collection of skin disease images and clinical notes. SkinGPT-4 demonstrated the ability to accurately diagnose various skin conditions and provide interactive treatment recommendations.[28] In addition to LLMs, AI-based methods, particularly those utilizing dermoscopic images, have shown promising results in assisting with melanoma detection. A systematic review by Patel et al. found that AI-based algorithms achieved higher ROC (>80%) compared to dermatologists in the detection of melanoma using dermoscopic images.[29] However, it is important to recognize that multi-modal LLMs are not yet reliable for independent clinical use. Their performance may be influenced by factors such as dataset limitations, image quality, and the lack of clinical context.

Despite these limitations, multi-modal LLMs may hold promise for applications in medical education. Sorin et al. explored the potential of multi-modal LLMs in ophthalmology education, suggesting that they could significantly impact this field by providing detailed explanations of ocular examination and imaging findings.[30] Similarly, in the context of melanoma and dermatology, multi-modal LLMs could assist students in identifying and describing lesion characteristics, considering differential diagnoses, and developing their clinical reasoning skills.

Mu et al. investigated the use of LLMs for management advice and found that ChatGPT provided more reliable and evidence-based recommendations compared to BARD and BingAI. However, all models were limited by a lack of depth and specificity, reducing their utility in individualized clinical decision-making.[19] This finding emphasizes the need for further refinement and validation of LLMs to ensure their recommendations align with clinical guidelines.

The limitations of this review include the small number of studies, heterogeneity in methodologies, and variations in evaluation metrics. Additionally, most studies had small sample sizes and did not involve patients in the question selection process. Furthermore, most studies focused on general melanoma questions rather than specific clinical scenarios.

In conclusion, this review highlights the potential of LLMs, particularly multi-modal models, in improving melanoma care through patient education, diagnosis, and management advice. Despite promising results, current LLM applications require further refinement to ensure clinical utility. Future studies should explore fine-tuning these models on large dermatological databases and incorporate expert knowledge.

Figures:

Figure 1. Hierarchy diagram of artificial intelligence (AI) terms.

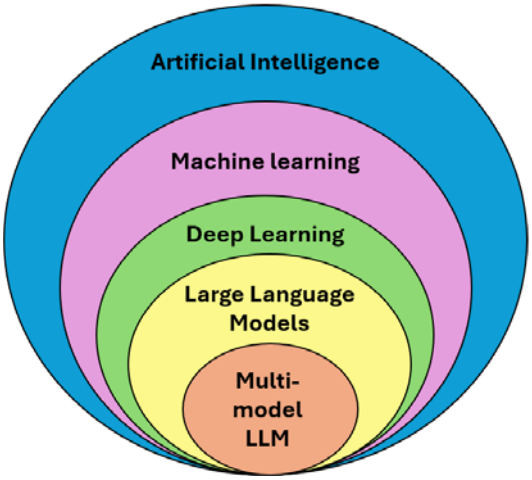


Figure 2. Applications of multi-modal LLMs in dermatology

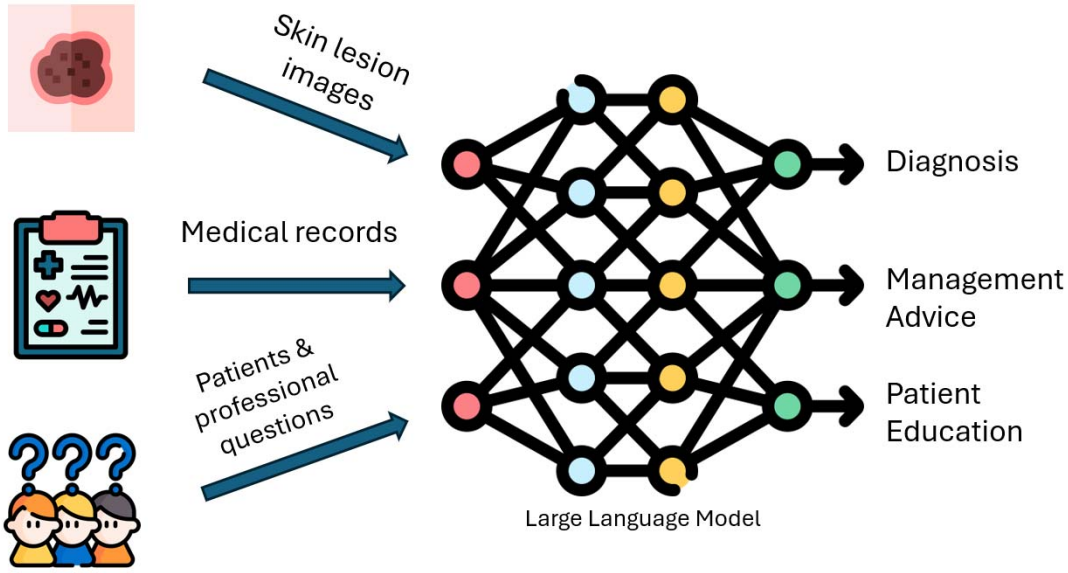
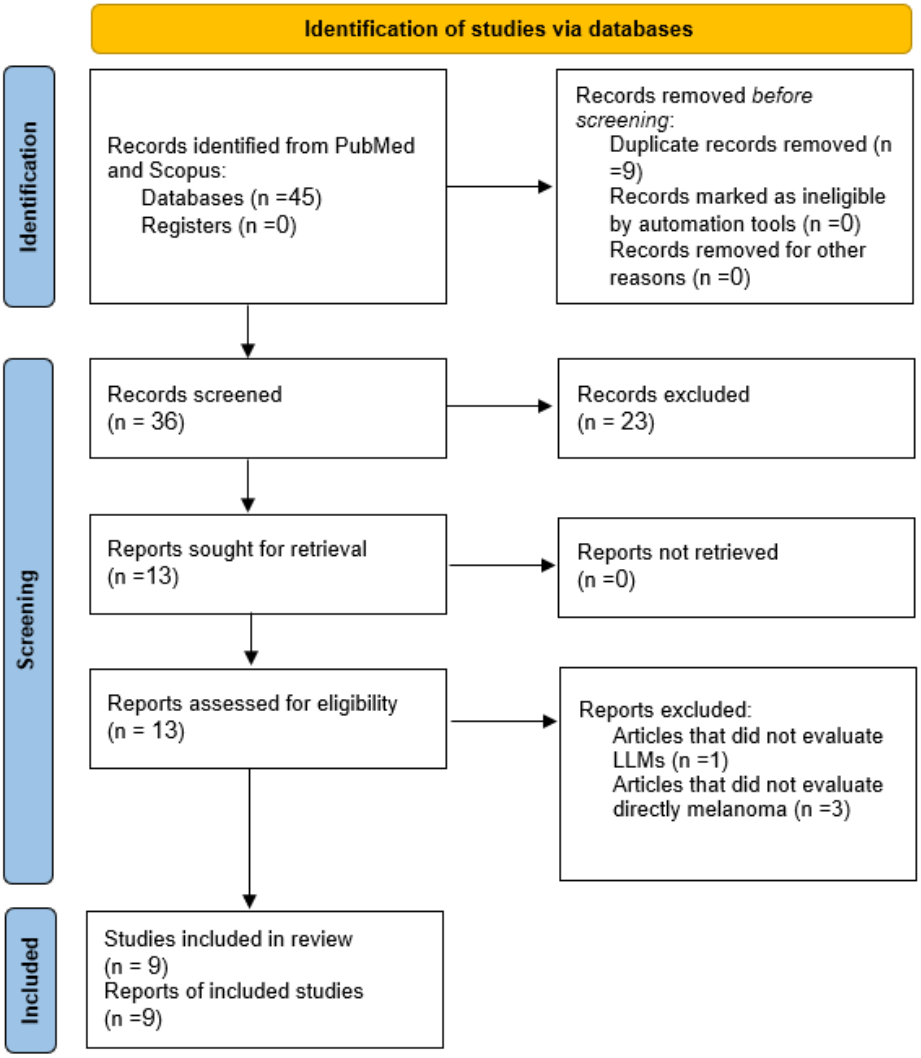


Figure 3. Flow Diagram of the Inclusion Process. Flow diagram of the search and inclusion process based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.



Tables:

Table 1. Details about reviewed articles

Group	Title	First Author	Journal	Year
Patient education	The utility of ChatGPT in generating patient-facing and clinical responses for melanoma	Jade N. Young[27]	Journal of the American Academy of Dermatology	2023
	Assessing large language models' accuracy in providing patient support for choroidal melanoma	Anguita R.[26]	Eye (Lond)	2024
	Comparative performance analysis of ChatGPT 3.5, ChatGPT 4.0 and Bard in answering common patient questions on melanoma	Deliyannis EP.[22]	Clinical and Experimental Dermatology	2024
	Readability and Health Literacy Scores for ChatGPT-Generated Dermatology Public Education Materials: Cross-Sectional Analysis of Sunscreen and Melanoma Questions	Roster K.[20]	JMIR Dermatology	2024
Melanoma Diagnosis	Assessing the Utility of Multimodal Large Language Models (GPT-4 Vision and Large Language and Vision Assistant) in Identifying Melanoma Across Different Skin Tones	Cirone K.[25]	JMIR Dermatology	2024
	Can ChatGPT Vision Diagnose Melanoma? An Exploratory	Shifai N.[23]	Journal of the American Academy of Dermatology	2024

	Diagnostic Accuracy Study			
	Evaluation of Vision LLMs GTP-4V and LLaVA for the Recognition of Features Characteristic of Melanoma	Akrout M.[21]	Journal of Cutaneous Medicine and Surgery	2024
Diagnosis of melanoma and medical education	Can Artificial Intelligence “Hold” a Dermoscope? The Evaluation of an Artificial Intelligence Chatbot to Translate the Dermoscopic Language	Karampinis E.[24]	Diagnostics (Basel)	2024
Management advice	Comparison of large language models in management advice for melanoma: Google's AI BARD, BingAI and ChatGPT	Mu X.[19]	Skin Health	2023

Table 2. A Summary of the reviewed articles

First Author	Model used	Objective	Reference Standard	Sample size	Main Findings	Conclusion
Patient education						
Jade N. Young	ChatGPT 4.0	Assess the appropriateness, clinical applicability, accuracy, and readability of ChatGPT 4.0 responses to	Three board-certified dermatologists	25 melanoma-related patient questions	**Accuracy**: (4.88/5) with agreement (80%, Fleiss K coefficient 0.808, P < .001). **Appropriateness**: 92% **Sufficiency**: 64% **Readability**: Average	ChatGPT 4.0 generates mostly accurate, but not sufficient, responses to melanoma patient questions, but it presents it at a level to

		melanoma-related questions			FRES 42.67(college-level readability)	advanced for the public use.
Anguita R.	ChatGPT 3.5, Bing AI, DocsGPT beta	Evaluate the accuracy of information provided by LLMs in response to common questions about choroidal melanoma.	Three ocular oncology experts	27 questions- 12 medical advice and 15 pre and post-operative advice	medical advice questions: **Accuracy**: GPT 3.5 92%, Bing AI 58%, DocsGPT 58% pre and post-operative advice: **Accuracy**: GPT 3.5 86%, Bing AI 86%, DocsGPT 73% 57% of responses **varied** across triplicated queries (Cohen's kappa =0.43, p < 0.05)	The three models demonstrate accuracy in response to most patient questions. There are no significant differences between the models.
Deliyannis EP.	ChatGPT 3.5, ChatGPT 4.0, Google Bard	Evaluate and compare the accuracy, readability, comprehensiveness, and reproducibility of responses provided by ChatGPT 3.5, ChatGPT 4.0, and Google Bard to common melanoma patient questions.	A consultant dermatologist and a senior dermatology trainee	205 questions were identified. **22** questions were selected	**Total score** for all 4 parameters), readability, comprehensiveness, reproducibility (out of 5): -ChatGPT 3.5: **4.51**, 4.68, 4.38, 4.41 -ChatGPT 4.0: **4.43**, 4.65, 4.4, 4.2 -Bard: **4.14,** 4.35, 4.09, 3.89 ChatGPT 3.5 and 4.0 **consistently** scored higher	ChatGPT and BARD may generate educate responses to common patient queries. Both versions of ChatGPT outperform BARD.

					than Bard for all parameters.	
Roster K.	ChatGPT	Evaluate the readability of ChatGPT generated dermatology public education materials on sunscreen and melanoma, and to determine if strategic prompting can improve readability to meet the American Medical Association (AMA) guidelines (6th grade reading level or less).	Readability was compared to AAD. Accuracy was evaluated by three dermatology residents. The study evaluated initial ChatGPT responses and responses after two rounds of strategic prompting.	**42** prompts were utilized, sourced from the American Academy of Dermatology (AAD) website's frequently asked questions (FAQs).	**Melanoma FAQs Readability:** (FRES score, average grade) AAD: 56.2, 9th grade ChatGPT initial: 46.5,10th grade ChatGPT with 2 prompt: 58.9, 8th grade -ChatGPT with 3 prompts: 59.3, 7th grade Prompting lowered the reading level vs. AAD (for 3 prompts P=.007) **Melanoma FAQs accuracy:** (scale from 1 to 3) -AAD: 2.82 -ChatGPT initial: 2.89 -ChatGPT with 2 prompt: 2.63 -ChatGPT with 3 prompts: 2.62	Using strategic prompting, ChatGPT could be used to enhance readability of medical data for melanoma patients. This prompting may result in less accuracy.
Melanoma Diagnosis						

Cirone K.	GPT-4V, LLaVA	Assess the ability of LLMs, specifically GPT-4 Vision and LLaVA, to accurately recognize and differentiate between melanoma and benign melanocytic nevi across different skin tones	Macroscopic images of melanoma and melanocytic nevi obtained from the MClass-D dataset.	20 text-based prompts, each tested on 3 images, resulting in **60** unique image-prompt combinations.	**GPT-4V Performance:** -Overall **accuracy**: 85% -Consistently provided descriptions of relevant ABCDE features. -Accurately identified melanoma across different skin tones and -recognized alterations in images. **LLaVA Performance:** -Overall **accuracy**: 45% -Unable to confidently identifying melanoma in individuals with darker skin tones -Vulnerable to visual prompt injection and manipulation, leading to diagnostic errors.	GPT-4V and LLaVA show potential in identifying melanoma across different skin tones, but further refinement is needed. GPT-4V outperforms LLaVA in overall accuracy.
Shifai N.	ChatGPT Vision	Assess the diagnostic accuracy of ChatGPT Vision in identifying melanoma using dermoscopic images	Dermoscopy images from ISIC archives.	100 melanocytic lesions (50 melanomas and 50 benign nevi)	The model provided 3 ranked differential diagnosis. **Top Diagnosis:** Sensitivity: 32% Specificity: 40% Diagnostic accuracy: 36%	ChatGPT Vision's current capabilities are inadequate for reliable melanoma diagnosis.

					Top-3 Differential Diagnoses: Sensitivity: 56% Specificity: 53.3% Diagnostic accuracy: 54.7% **Malignant vs. Benign (Top Diagnosis):** Sensitivity: 46% Specificity: 78% Diagnostic accuracy: 62% **Malignant vs. Benign (Top-3 Diagnoses):** Sensitivity: 78% Specificity: 46.7% Diagnostic accuracy: 62.3%	
Akrout M.	GTP-4V, LLaVA	Assess the ability of vision LLMs to recognize, classify, and appropriately comment on the ABCDE features of melanoma lesions.	Macroscopic images obtained from the publicly available MD-class dataset and Dermnet NZ	55 unique text-based prompts consisting of questions and instructions, and image-based prompts	**GTP-4V Performance:** -Accurately described asymmetry, border, color, diameter, and evolution. - Inconsistently identified melanoma subtypes -Vulnerable to **visual prompt** injections**.**	GTP-4V outperformed LLaVA. While GTP-4V and LLaVA show promise in recognizing features characteristic of melanoma, both models require further refinement to improve diagnostic accuracy

				highlighting areas of focus	**LLaVA Performance:** **-** Accurately described asymmetry, border, and color**.** **-** inaccurately assessed diameter and evolution. **-** Inconsistently identified melanoma subtypes -Less vulnerable to **visual prompt injections**	and consistency.
Karampinis E.	ChatGPT 3.5	Assess the clarity of dermoscopic language translated by an AI chatbot and its role in facilitating accurate diagnoses and educational opportunities for novice dermatologists	30 participants with a certification in dernoscopy	The survey comprised instances of dermoscopic descriptions, including **3 pigmented lesions** (1 melanoma and 2 nevi)	**pigmented lesions scores:** (scale 1 to 3) completeness:2.4 ± 0.88 Helpful to diagnosis:2.8 ± 0.48 Teaching tool: 2.7 ± 0.59 For pigmented lesions, incorporating clinical patient data did not significantly change the results.	AI chatbot demonstrates potential in translating dermoscopic language but requires further development to improve its accuracy and reliability for clinical use
Manageme nt advice						
Mu X.	ChatGPT -4,	Compare the performance of	2 plastic surgent	5 questions on	**readability**:	ChatGPT provides more reliable,

BingAI,Google's AI BARD	Google's AI BARD, BingAI, and ChatGPT-4 in providing melanoma management advice based on current clinical guidelines and literature.	residents, 1 registrar and 3 specialist plastic surgeons	melanoma management	(Flesch Reading Ease Score, Flesch-Kincaid Grade Level) ChatGPT: 35.42, 11.98 BARD: 32.1 , 15.03 BingAI: 29.88 ,13.58 the mean readability exhibited considerable similarity. **reliability :** DISCERN score : ChatGPT 58 (+-6.44) BARD 36.2 (+-34.06) BingAI's 49.8 (+-22.28). The only statistically significant test was comparing ChatGPT to BARD for the DISCERN score(p-value 0.04)	evidence-based clinical advice than BARD and BingAI. However, all models lack depth and specificity, limiting its use in individualized clinical decision-making.

Table 3: Advantages and Challenges of the reviewed articles

First Author	Advantages	Challenges
Patient education		
Jade N. Young	1. The responses were evaluated by three board-certified dermatologists, ensuring that the	1. Patients were not involved in the question selection process, potentially missing out on patient

	assessment of the AI's performance was thorough and conducted by knowledgeable professionals 2. The agreement between the evaluators was statistically significant	perspectives
Anguita R.	1. The study compares 3 different LLMs, offering a broad perspective on their performance 2. The study relies on the assessment of three experts who were blinded to the LLM they were using	1. The study is limited to a subtype of melanoma 2. The study focussed only on accuracy and did not evaluate other aspects
Deliyannis EP.	1. Questions were identified from online sources such as Facebook groups, national foundations, and charity websites, increasing the relevance and practical importance of the questions evaluated 2. The study compares three different LLMs, offering a broad perspective on their performance 3. The responses were assessed for accuracy, readability, comprehensiveness, and reproducibility, providing a thorough evaluation	1. Only 2 assessors were involved in scoring the responses, which might limit the robustness of the evaluation 2. Readability was not assessed using FRES score
Roster K.	1. The use of multiple readability and health literacy tools provides a thorough evaluation of the text readability 2. Accuracy was assessed by 3 dermatology residents, ensuring the reliability of the content evaluation 3. The use of multiple prompts on the same FAQ demonstrates the model's strength in improving	1. The study only evaluates ChatGPT, limiting the comparison with other LLMs 2. It is unclear how many prompts specifically addressed melanoma

	readability	
Diagnosis of melanoma		
Cirone K.	1. The use of Multiple LLMs offers a broad perspective on their performance. 2. Evaluation of the models' ability to handle image manipulations and consider skin tone variations demonstrates the models' effectiveness across different diagnostic factors	1. Absence of statistical significance tests 2. The number of benign nevi vs. melanomas that were recognized or un-recognized is not specified. Thus, the reader cannot interpret the sensitivity and specificity of the diagnosis 3. the study does not specify the number of evaluators who assessed the accuracy of the results, as well as the unknown proficiency of these evaluators
Shifai N.	1. The study uses a balanced data set with an equal number of melanomas and benign nevi, thus improving the credibility of the study 2. The evaluation uses sensitivity and specificity metrics to assess the model's diagnostic performance for both positive and negative cases	1. The absence of intermediate melanocytic lesions, such as dysplastic nevi, oversimplifies the evaluation compared to routine clinical settings 2. Factors such as anatomic site, skin type, nevi subtype, melanoma subtype, and tumour thickness were not considered in the analyses
Akrout M.	1. The study utilized a balanced data set covering various melanoma stages which enhances the robustness of the evaluation 2. The evaluation included metrics for describing ABCDE features, identifying melanoma subtypes, and handling visual prompt injections, offering a detailed assessment of model performance	1. No statistical tools were used 2. The study utilized "textbook" or idealized images of melanoma, which may not accurately represent the diverse range of lesions encountered in real-world clinical settings. 3. The evaluators' identities and their proficiency in interpreting the model outcomes are unknown

	3. The Use of Multiple LLMs offers a broad perspective on their performance	
Karampinis E.	1. The results are based on feedback from 30 participants, providing diverse insights into the chatbot's performance 2. The prompts were evaluated both with and without incorporating additional clinical patient data	1. Only three descriptions of pigmented lesions were used 2. The study did not focus specifically on melanotic lesions
Management advice		
Mu X.	1. The study involves a panel of experienced board-certified plastic surgeons to assess the responses 2. The use of multiple readability matrixes provides a thorough evaluation of the text readability 3. The comparison of Multiple LLMs offers a broad perspective on their performance	1. The small number of questions limits the generalizability of the results 2. The questions examined were mostly general and did not address to a patient's clinical background. 3. The study evaluates LLMs' responses based solely on existing guidelines, without considering newer research that may provide more up-to-date information

References

1. Usman Hadi, M. *et al.* Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects Large Language Models: A Comprehensive Survey of Applications, Challenges, Limitations, and Future Prospects. doi:10.36227/techrxiv.23589741.v4.
2. Clusmann, J. *et al.* The future landscape of large language models in medicine. doi:10.1038/s43856-023-00370-1.

3. Mudrik, A. *et al.* Exploring the role of Large Language Models (LLMs) in hematology: a systematic review of applications, benefits, and limitations. *medRxiv* 2024.04.26.24306358 (2024) doi:10.1101/2024.04.26.24306358.

4. Preiksaitis, C. *et al.* The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review. *JMIR Med Inform* **12**, e53787 (2024).

5. Pressman, S. M. *et al.* Clinical and Surgical Applications of Large Language Models: A Systematic Review. *Journal of Clinical Medicine 2024, Vol. 13, Page 3041* **13**, 3041 (2024).

6. Klang, E., Sourosh, A. & Nadkarni, G. N. Evaluating the role of ChatGPT in gastroenterology: a comprehensive systematic review of applications, benefits, and limitations. *Therap Adv Gastroenterol* **16**, (2023).

7. Glicksberg, B. S. *et al.* Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room. *J Am Med Inform Assoc* (2024) doi:10.1093/JAMIA/OCAE103.

8. Sallam, M., Salim, N. A., Barakat, M. & Al-Tammemi, A. B. ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. *Narra J* **3**, e103–e103 (2023).

9. Deng, J., Heybati, K. & Shammas-Toma, M. When vision meets reality: Exploring the clinical applicability of GPT-4 with vision. *Clin Imaging* **108**, 110101 (2024).

10. Arias-Santiago, S. *et al.* ChatGPT in dermatology: exploring the limited utility amidst the tech hype. (2024) doi:10.3389/fmed.2023.1308229.

11. Sarker, I. H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. **2**, 420 (2021).

12. Sheikh, H., Prins, C. & Schrijvers, E. Artificial Intelligence: Definition and Background. 15–41 (2023) doi:10.1007/978-3-031-21448-6_2.

13. Almarie, B., Teixeira, P. E. P., Pacheco-Barrios, K., Rossetti, C. A. & Fregni, F. Editorial – The Use of Large Language Models in Science: Opportunities and Challenges. *Princ Pract Clin Res* **9**, 1 (2023).

14. Nazi, Z. Al & Peng, W. Large Language Models in Healthcare and Medical Domain: A Review. *Informatics 2024, Vol. 11, Page 57* **11**, 57 (2024).

15. Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, (2021).

16. Moons, K. G. M. *et al.* Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* **11**, (2014).

17. Schiavo, J. H. PROSPERO: An International Register of Systematic Review Protocols. *Med Ref Serv Q* **38**, 171–180 (2019).

18. Whiting, P. F. *et al.* QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* **155**, 529–536 (2011).

19. Mu, X. *et al.* Comparison of large language models in management advice for melanoma: Google's AI BARD, BingAI and ChatGPT. (2023) doi:10.1002/ski2.313.

20. Roster, K. *et al.* Readability and Health Literacy Scores for ChatGPT-Generated Dermatology Public Education Materials: Cross-Sectional Analysis of Sunscreen and Melanoma Questions. *JMIR Dermatol* **7**, (2024).

21. Akrout, M., Cirone, K. D. & Vender, R. Evaluation of Vision LLMs GTP-4V and LLaVA for the Recognition of Features Characteristic of Melanoma. *J Cutan Med Surg* **28**, 98–99 (2024).

22. Deliyannis, E. P., Paul, N., Patel, P. U. & Papanikolaou, M. Comparative performance analysis of ChatGPT 3.5, ChatGPT 4.0 and Bard in answering common patient questions on melanoma. *Clin Exp Dermatol* **49**, 743–746 (2024).

23. Shifai, N., van Doorn, R., Malvehy, J. & Sangers, T. E. Can ChatGPT vision diagnose melanoma? An exploratory diagnostic accuracy study. *J Am Acad Dermatol* **90**, 1057–1059 (2024).

24. Karampinis, E. *et al.* Can Artificial Intelligence 'Hold' a Dermoscope?-The Evaluation of an Artificial Intelligence Chatbot to Translate the Dermoscopic Language. *Diagnostics (Basel)* **14**, (2024).

25. Cirone, K., Akrout, M., Abid, L. & Oakley, A. Assessing the Utility of Multimodal Large Language Models (GPT-4 Vision and Large Language and Vision Assistant) in Identifying Melanoma Across Different Skin Tones. *JMIR Dermatol* **7**, (2024).

26. Anguita, R., Downie, C., Ferro Desideri, L. & Sagoo, M. S. Assessing large language models' accuracy in providing patient support for choroidal melanoma. *Eye (Lond)* (2024) doi:10.1038/S41433-024-03231-W.

27. Young, J. N. *et al.* The utility of ChatGPT in generating patient-facing and clinical responses for melanoma. *J Am Acad Dermatol* **89**, 602–604 (2023).

28. Zhou, J. *et al.* Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nat Commun* **15**, (2024).

29. Patel, R. H., Foltz, E. A., Witkowski, A. & Ludzik, J. Analysis of Artificial Intelligence-Based Approaches Applied to Non-Invasive Imaging for Early Detection of Melanoma: A Systematic Review. *Cancers (Basel)* **15**, (2023).

30. Sorin, V. *et al.* GPT-4 Multimodal Analysis on Ophthalmology Clinical Cases Including Text and Images. *medRxiv* 2023.11.24.23298953 (2023) doi:10.1101/2023.11.24.23298953.