This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

DermAgent: A Multi-Tool Agentic System for Evidence-Grounded and Traceable Reasoning in Dermatology

Yize Liu[1,2][0009−0001−3831−8116], Siyuan Yan[1,2][0000−0002−6372−8336]⋆, Ming Hu[1,2], Lie Ju[3][0000−0002−6687−7054], Xieji Li[1,2], Feilong Tang[1,2], Wei Feng[1,2], and Zongyuan Ge[1,2][0000−0002−5880−8673]

[1] AIM for Health Lab, Faculty of Information Technology, Monash University, Melbourne, Australia
[2] Faculty of Information Technology, Monash University, Melbourne, Australia
siyuan.yan@monash.edu
[3] University College London, Institute of Ophthalmology, London, United Kingdom

Abstract. Dermatological diagnosis requires integrating fine-grained visual perception with expert clinical knowledge. Although Multimodal Large Language Models (MLLMs) facilitate interactive medical image analysis, their application in dermatology is hindered by insufficient domain-specific grounding and hallucinations. To address these issues, we propose DermAgent, a collaborative multi-tool agent that orchestrates seven specialized vision and language modules within a Plan–Execute–Reflect framework. DermAgent delivers stepwise, traceable diagnostic reasoning through three core components. First, it employs complementary visual perception tools for comprehensive morphological description, dermoscopic concept annotation, and disease diagnosis. Second, to overcome the lack of domain prior, a dual-modality retrieval module anchors every prediction in external evidence by cross-referencing 413,210 diagnosed image cases and 3,199 clinical guideline chunks. To further mitigate hallucinations, a deterministic critic module conducts strict post-hoc auditing via confidence, coverage, and conflict gates, automatically detecting inter-source disagreements to trigger targeted self-correction. Extensive experiments on five dermatology benchmarks demonstrate that DermAgent consistently outperforms state-of-the-art MLLMs and medical agent baselines across zero-shot fine-grained disease diagnosis, concept annotation, and clinical captioning tasks, exceeding GPT-4o by 17.6% in skin disease diagnostic accuracy and 3.15% in captioning ROUGE-L. Our code is available at https://github.com/YizeezLiu/DermAgent.

Keywords: Agentic AI · Dermatology · Medical Image Analysis

1 Introduction

Accurate dermatological diagnosis is a complex process that demands far more than visual pattern recognition [18]. With a vast and long-tailed taxonomy of

⋆ Corresponding author: Siyuan Yan (siyuan.yan@monash.edu).

skin conditions [6, 7], clinicians must integrate subtle morphological cues with ontological knowledge of diagnostic mimics and rare variants. This workflow is inherently multimodal and iterative: it involves formulating differential diagnoses, verifying hypotheses against dermoscopic findings, and cross-referencing these findings with medical literature. Developing automated systems that can replicate not only the outcome of expert diagnosis but also the evidence-based reasoning underlying it remains an open challenge.

Deep learning has achieved performance comparable to that of dermatologists on static classification tasks [6, 9], and recent foundation models [30, 15] have further advanced visual representation learning. However, these models operate primarily as isolated, task-specific systems. This lack of flexibility restricts their utility as interactive diagnostic aids [11]. To enable interactive reasoning over diverse clinical queries, researchers have turned to Multimodal Large Language Models (MLLMs) [22, 3, 17, 4]. While these systems integrate visual understanding with natural language generation to facilitate open-ended dialogue, their clinical applicability is hindered by two fundamental limitations. First, their reasoning is grounded solely in parametric knowledge acquired during pretraining, with no mechanism to retrieve or verify information against external clinical references such as diagnostic guidelines or curated case repositories. This absence of evidence-based grounding makes them prone to hallucinations [19], a particularly critical risk in dermatology where morphologically similar lesions demand precise differential diagnosis [10, 23]. Second, accurate dermatological diagnosis requires integrating signals across diverse examination dimensions. This process encompasses visual pattern classification to map images to candidate diagnoses, structured dermoscopic concept extraction for detecting standardized features (e.g., pigment networks and border irregularity), and free-form morphological description in natural language. Existing MLLMs lack both the specialized perception models needed to capture these domain-specific features and the reasoning mechanisms to detect and resolve the frequent contradictions.

Agentic AI provides a promising approach to address these challenges by orchestrating specialized external tools [20, 33, 8]. Motivated by this paradigm, we propose **DermAgent**, a multi-tool collaborative agent designed for comprehensive and interpretable dermatological image analysis. First, to overcome the limited transparency of end-to-end inference in isolated models, DermAgent employs an LLM controller that operates through a Plan–Execute–Reflect loop. This controller decomposes complex clinical queries into structured sub-tasks and dispatches them to specialized vision and language tools, thereby providing stepwise reasoning where intermediate outputs are individually inspectable. Second, to prevent reliance on parametric memory alone and to anchor each reasoning step in verifiable clinical evidence, we equip the system with a dual-modality knowledge retrieval module. This module comprises Case-Based Image Retrieval (Case RAG) and Guideline-Grounded Text Retrieval (Guideline RAG), which provide complementary knowledge sources that cross-validate the visual findings of the agent. Third, to detect and resolve inter-source conflicts and flag low-confidence outputs that arise across these diverse sources, we de-

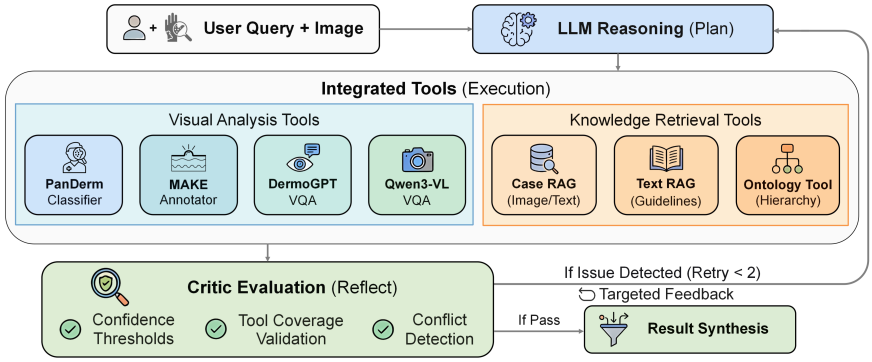

Fig. 1. Overview of the proposed DermAgent framework. An LLM controller orchestrates specialized visual perception and knowledge retrieval tools via an iterative Plan–Execute–Reflect loop. A deterministic Critic module further audits the accumulated evidence chain to trigger targeted self-correction.

sign a deterministic Critic module. This module performs post-hoc auditing of the assembled evidence chain and directs the agent to collect targeted external validation before finalizing the diagnosis.

Our contributions are as follows: (1) We introduce DermAgent, a collaborative agent that orchestrates seven specialist vision and language tools, providing stepwise, inspectable reasoning. (2) We design a dual-modality retrieval module combining Case RAG over 413,210 diagnosed cases with Guideline RAG over 3,199 clinical guideline chunks, grounding predictions in traceable external knowledge. (3) We design a deterministic Critic module with confidence, coverage, and conflict gates that trigger targeted self-correction to suppress hallucinations. (4) Extensive experiments on five datasets demonstrate consistent improvements over MLLM and agent baselines across disease diagnosis, concept annotation, and clinical captioning.

2 DermAgent

We formulate dermatological diagnosis as a multi-step evidence reasoning problem. Given a query tuple (I, q) consisting of a skin image I and a clinical question q, the goal is to generate a response R accompanied by a traceable evidence chain $\mathcal{E}$, where each element $e_i \in \mathcal{E}$ represents a verified finding from a specialized tool. As illustrated in Fig. 1, DermAgent operates through a *Plan–Execute–Reflect* loop, dynamically orchestrating seven specialist tools to simulate the clinical workflow of observation, analysis, and reference consultation.

2.1 Architecture and Workflow

The agent architecture is modeled as a directed cyclic graph comprising three nodes: *Chatbot*, *Tools*, and *Critic*. The execution flow is formalized in Algo-

Algorithm 1 DermAgent Plan–Execute–Reflect Framework

Require: Q: User query, I: Dermatological image, $\mathcal{T}$: Specialist tool set, $k_{\max}$: Maximum retries
Ensure: R: Final response with evidence chain $\mathcal{E}$

$k \leftarrow 0, \quad \mathcal{E} \leftarrow \emptyset, \quad \mathcal{G} \leftarrow \emptyset$
$\mathcal{S} \leftarrow \textsc{AnalyzeTask}(Q, I)$ ▷ Identify task type and requirements
repeat
  — **Plan** —
  $\mathcal{P}_k \leftarrow \textsc{Plan}(\mathcal{S}, \mathcal{E}, \mathcal{G} \mid \mathcal{T})$ ▷ LLM selects tools and queries
  — **Execute** —
  for $(t_i, \theta_i) \in \mathcal{P}_k$ **do** ▷ θ_i: tool-specific query/parameters
    $e_i \leftarrow \textsc{Execute}(t_i, \theta_i, I)$
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{(t_i, \theta_i, e_i)\}$
  end for
  — **Reflect (Critic)** —
  $f_{\text{conf}} \leftarrow \textsc{CheckConfidence}(\mathcal{E})$ ▷ Low-confidence detection
  $f_{\text{cov}} \leftarrow \textsc{CheckCoverage}(\mathcal{S}, \mathcal{E})$ ▷ Missing tool detection
  $f_{\text{con}} \leftarrow \textsc{DetectConflicts}(\mathcal{E})$ ▷ Cross-tool disagreement
  if $(f_{\text{conf}} \vee f_{\text{cov}} \vee f_{\text{con}})$ **and** $k < k_{\max}$ **then**
    $\mathcal{G} \leftarrow \mathcal{G} \cup \textsc{Feedback}(f_{\text{conf}}, f_{\text{cov}}, f_{\text{con}}, \mathcal{E})$ ▷ Directional critique
    $k \leftarrow k + 1$
  end if
until $\neg(f_{\text{conf}} \vee f_{\text{cov}} \vee f_{\text{con}})$ **or** $k \geq k_{\max}$
— **Synthesis** —
$R \leftarrow \textsc{Synthesize}(\mathcal{E}, Q)$ ▷ Cross-validate and generate evidence-grounded answer
return R

rithm 1. The system maintains a dynamic evidence chain $\mathcal{E}$ (accumulating verified findings), a feedback memory $\mathcal{G}$ (from the Critic), and a retry counter k.

Task Analysis & Planning. Initially, the agent performs task analysis to determine the problem scope $\mathcal{S}$. In the planning phase of iteration k, the *Chatbot* node (powered by GPT-4o [22]) synthesizes the current evidence $\mathcal{E}$ and prior critic feedback $\mathcal{G}$ to generate a coherent plan $\mathcal{P}_k = \{(t_i, \theta_i)\}$, consisting of specific tool calls t_i and their parameters θ_i.

Execute & Reflect. The *Tools* node executes the planned calls in parallel, collecting results e_i and updating the evidence chain: $\mathcal{E} \leftarrow \mathcal{E} \cup \{(t_i, \theta_i, e_i)\}$. To prevent hallucinations, the *Critic* node acts as a quality controller. It evaluates the updated $\mathcal{E}$ against three logic checks (detailed in Sec. 2.3). If deficiencies are found (e.g., low confidence or missing coverage) and $k < k_{\max}$, the Critic generates directional feedback $\mathcal{G}$ and routes the flow back to the Chatbot for refinement. Otherwise, the agent proceeds to *Synthesis* to generate the final response R.

2.2 Specialist Tool Modules

DermAgent assembles seven specialist tools organized into two functional groups. *Visual perception modules* are adopted from established foundation models [30, 29, 24, 3], selected for their complementary representational axes and state-of-the-art performance on dermatological benchmarks. *Knowledge retrieval mechanisms* ground agent predictions in non-parametric external evidence at inference

time, providing verifiable, source-attributed reasoning that cannot be fabricated from model weights alone.

PanDerm Classifier. This module performs zero-shot skin disease classification using DermLIP [30] over a diverse disease taxonomy. Given an image I and a candidate label set $\mathcal{C} = \{c_1, \ldots, c_N\}$, DermLIP computes cosine similarity between the image embedding and the text embedding of each candidate, returning a ranked prediction list with calibrated confidence scores.

MAKE Concept Annotator. MAKE [29] detects dermoscopic concepts (e.g., pigment network, streaks) via zero-shot CLIP-based scoring. For an image I and feature set $\mathcal{F} = \{f_1, \ldots, f_M\}$, it derives visual–semantic alignments to output a structured annotation set $\mathcal{A} \subseteq \mathcal{F}$.

DermoGPT (Dermatology VQA). DermoGPT [24] is a dermatology specialized vision–language model that provides free-form visual question answering for morphological description and diagnostic reasoning.

Qwen3-VL (General VQA). We employ Qwen3-VL-8B-Instruct [3] specifically for complementary VQA tasks.

Case-Based Image Retrieval (Case RAG). The Case RAG module encodes a query image with the DermLIP encoder (shared with the PanDerm classifier) into a 512-dimensional embedding and performs cosine similarity search over a vector database containing 413,210 clinically diagnosed cases from Derm1M [28]. Each retrieved entry carries a disease label, a hierarchical diagnostic category, and a clinical description, providing non-parametric evidence grounded in real diagnostic records. To avoid test-set leakage, each image is assigned an anonymized image ID, and retrieval candidates with the same ID as the query image are strictly excluded during evaluation. Original file names and dataset paths are also hidden from the LLM orchestrator to prevent filename-level leakage.

Guideline-Grounded Text Retrieval (Guideline RAG). The Guideline RAG module queries a curated knowledge base of 3,199 document chunks compiled from DermNet [1] (2,751 chunks) and Mayo Clinic [2] (448 chunks) dermatology references. Retrieval follows a four-stage hybrid pipeline. The input query first undergoes domain-specific stop-word filtering to remove generic medical and interrogative terms. The filtered query is then encoded by Qwen3-Embedding-8B [32] into 4,096-dimensional embeddings for semantic vector search, while simultaneously tokenized for full-text keyword matching over indexed fields. The two candidate lists are merged via Reciprocal Rank Fusion (RRF, k=60) and re-ranked by Qwen3-Reranker-0.6B [32], a 0.6-billion-parameter cross-encoder that scores each query–document pair through a binary relevance prompt. The module returns disease names, clinical guideline sections, and source URLs, supplying guideline-grounded evidence for differential diagnosis and terminology verification.

Ontology Tool. This tool is a hierarchical knowledge graph of skin disease taxonomy with fuzzy name matching. Given a query mode $m \in \{\text{hierarchy}, \text{children}, \text{siblings}, \text{search}\}$ and a disease name d, it returns the corresponding set of disease entities or hierarchical relationships.

2.3 Critic-Driven Reflection

After each execution round, a deterministic Critic module evaluates the accumulated evidence chain $\mathcal{E}$ against three gate conditions (Algorithm 1, lines 11–14). If *any* condition is satisfied and $k < k_{\max}$, the Critic routes control back to the Chatbot node, where the LLM controller re-examines the image and current evidence to autonomously plan a new round of evidence collection. Otherwise, $\mathcal{E}$ is forwarded to final synthesis.

Confidence Check (f_{conf}). The Critic flags low-confidence outputs based on empirically set thresholds: PanDerm predictions with cosine similarity below 90%, or RAG retrievals with similarity below 80%. A flag is raised only when an uninvoked actionable tool remains available, ensuring retries introduce genuinely new evidence rather than repeating prior calls.

Coverage Check (f_{cov}). The Critic verifies that task-critical specialist tools have been invoked for the given query type. This prevents the LLM controller from short-circuiting to a VQA model alone without engaging task-specific modules (e.g., PanDerm for diagnosis, MAKE for concept annotation), ensuring each final answer is grounded in evidence from multiple complementary sources.

Conflict Check (f_{con}). The Critic compares top-ranked predictions across visual specialist tools (PanDerm and Case RAG) and flags unresolved disagreements. Conflicts already addressed through prior Guideline RAG retrieval are accepted, preventing unproductive retry loops.

3 Experiments

Experimental Setup. DermAgent is evaluated on five established dermatology benchmarks spanning three complementary tasks. For zero-shot classification: HAM10000 [26] (7 diseases, 642 images) and SNU [12] (134 fine-grained disease classes, 500 images). For dermoscopic concept annotation: Derm7pt [14] (7 dermoscopic concepts, 100 images) and SkinCon [5] (32 clinical concepts, 100 images). For clinical captioning: SkinCAP [25] (100 images). We employ stratified sampling across all benchmarks and oversample underrepresented classes to ensure balanced evaluation coverage.

3.1 Main Results

Table 1 presents performance comparisons across all benchmarks. In zero-shot classification, DermAgent achieves 61.83% accuracy on HAM10000, surpassing MedAgent-Pro (57.63%) by +4.20% and the strongest MLLM baseline (Hulu-Med-7B, 52.65%) by +9.18%. On the fine-grained SNU dataset (134 classes), DermAgent attains 32.60%, achieving more than twice the accuracy of the next best result (GPT-4o, 15.00%). For concept annotation, DermAgent achieves the highest F1-Macro on both Derm7pt (65.06%, versus 56.86% for DermoGPT-RL, +8.20%) and SkinCon (32.95%, versus 29.56% for GPT-4o, +3.39%), confirming that cross-validation between structured annotations and visual descriptions

Table 1. Performance comparison on dermatological diagnosis (Accuracy), concept annotation (F1-Macro), and captioning (ROUGE-L) tasks (%).

Model	Type	Diagnosis		Concept Annotation		Captioning
		HAM10000	SNU	Derm7pt	SkinCon	SkinCAP
LLaVA-Med-v1.5 [17]	Medical MLLM	44.24	1.20	51.70	13.10	15.32
HuatuoGPT [4]	Medical MLLM	51.40	4.00	53.43	9.49	14.32
Hulu-Med-7B [13]	Medical MLLM	52.65	0.80	51.63	9.63	11.43
DermoGPT-RL [24]	Dermatology MLLM	50.00	9.20	56.86	20.72	15.41
SkinVL-PubMM [31]	Dermatology MLLM	45.17	3.40	53.14	13.20	14.44
Qwen3-VL-8B [3]	General MLLM	51.09	7.80	53.70	22.82	12.47
GPT-4o [22]	General MLLM	48.91	15.00	54.14	29.56	16.33
GPT-5.2 [21]	General MLLM	35.98	14.80	53.86	26.62	12.35
MDAgents [16]	Medical Agent	16.82	11.40	36.14	23.93	11.99
MedAgent-Pro [27]	Medical Agent	57.63	11.60	64.82	18.34	11.48
DermAgent (Ours)	**Medical Agent**	**61.83**	**32.60**	**65.06**	**32.95**	**19.48**

Table 2. Ablation study on SkinCAP clinical captioning (%).

Configuration	ROUGE-L	Δ
Full Agent (w/ Critic)	**19.48**	**+2.21**
Full Agent (w/o Critic)	17.27	—
w/o Case RAG	15.80	−1.47
w/o Guideline RAG	16.28	−0.99
w/o DermoGPT	16.72	−0.55
w/o PanDerm	16.76	−0.51
w/o MAKE	16.79	−0.48
w/o Ontology	17.12	−0.15

effectively mitigates false positives. In clinical captioning, DermAgent obtains a ROUGE-L of 19.48% on SkinCAP, outperforming GPT-4o (16.33%) by +3.15%.

To quantify the contribution of individual components, we conduct a leave-one-out (LOO) ablation study on SkinCAP (Table 2).

Critic mechanism. The Critic provides the largest improvement (+2.21%), confirming that automatic consistency checking and targeted retries yield more coherent and evidence-grounded captions.

Retrieval modules. Among individual tools, Case RAG ($\Delta = -1.47\%$) and Guideline RAG ($\Delta = -0.99\%$) are the two most impactful components. Case RAG grounds the clinical impression in real diagnostic records, while Guideline RAG supplies precise medical terminology that improves lexical alignment with reference descriptions.

Visual perception and ontology tools. Removing DermoGPT ($\Delta = -0.55\%$) or PanDerm ($\Delta = -0.51\%$) moderately degrades performance, indicating their complementary roles in providing morphological descriptions and diagnostic hypotheses. MAKE ($\Delta = -0.48\%$) contributes structured dermoscopic features, while the ontology tool has minimal impact on captioning.

3.2 Qualitative Analysis

	**Ground Truth:** Granuloma annulare is a relatively rare skin lesion characterized by localized red patches with raised ring-shaped edges and sometimes crusting. This condition may be caused by allergic reactions, infections, or other inflammatory processes [. . .]
GPT-4o (Baseline)	The skin lesion appears to be a circular patch located on the hand [. . .] The borders are distinct but not sharply demarcated [. . .] These features are often associated with conditions like tinea corporis (ringworm) or other types of dermatitis.
DermAgent (Ours)	**Round 1:** DermoGPT → *erythematous, raised borders, dorsal hand* → Eczema; PanDerm → Eczema (100%); Case RAG → Granuloma Annulare (×4, sim>0.73); Guideline RAG → eczema features ***Critic:*** f_{con} *conflict* → *image reinject* → *LLM observes round shape, central clearing, raised borders* → *conflicts with Eczema* **Round 2:** PanDerm (refined) → Granuloma Annulare (100%); Guideline RAG → GA clinical features & guidelines The image shows a round, well-defined lesion on the dorsum of the hand with erythematous, slightly raised borders and central clearing, consistent with granuloma annulare. Granuloma annulare is a benign inflammatory skin condition [. . .]

Fig. 2. Qualitative comparison on a representative case in the captioning task. Green highlights correct descriptions; red highlights hallucinated diagnoses.

Figure 2 presents a representative case of granuloma annulare. While GPT-4o misidentifies the lesion as tinea corporis, DermAgent overcomes this bias toward common classes through critic-driven self-correction. Initially, DermoGPT accurately describes the morphological features of the lesion but hallucinates a diagnosis of eczema. PanDerm similarly predicts eczema. However, the Case RAG module retrieves four highly similar cases of granuloma annulare. The f_{con} gate detects this discrepancy and triggers a retry mechanism with image reinjection. Upon re-evaluating the visual evidence, the LLM controller recognizes that the observed raised borders and central clearing conflict with the diagnostic criteria for eczema. Subsequently, a refined query to PanDerm confirms the diagnosis of granuloma annulare, while the Guideline RAG module retrieves the appropriate clinical guidelines. This process provides the precise terminology required to synthesize an accurate and evidence-grounded caption.

4 Conclusion

We introduced DermAgent, a collaborative multi-tool agent for comprehensive dermatological image analysis. It orchestrates specialized visual perception mod-

ules and complementary Case RAG and Guideline RAG within a Plan–Execute–Reflect framework, bridging data-driven pattern recognition and knowledge-grounded diagnosis. Evaluations across five benchmarks show that DermAgent outperforms state-of-the-art MLLMs and agents in diagnosis, concept annotation, and captioning, with ablation studies confirming the critic-driven self-correction mechanism and retrieval modules as the most critical components.

Acknowledgments. This work was supported by the Australian Centre of Excellence in Melanoma Imaging and Diagnosis (ACEMID).

Disclosure of Interests. The authors declare no competing interests.

References

1. DermNet. https://dermnetnz.org/
2. Mayo Clinic - Medical Diseases & Conditions. https://www.mayoclinic.org/diseases-conditions
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X. et al.: Qwen3-VL Technical Report (Nov 2025). https://doi.org/10.48550/arXiv.2511.21631
4. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S. et al.: HuatuoGPT-Vision, Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale (Sep 2024). https://doi.org/10.48550/arXiv.2406.19280
5. Daneshjou, R., Yuksekgonul, M., Cai, Z.R., Novoa, R., Zou, J.: SkinCon: A skin disease dataset densely annotated by domain experts for fine-grained model debugging and analysis (Feb 2023). https://doi.org/10.48550/arXiv.2302.00785
6. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M. et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (Feb 2017). https://doi.org/10.1038/nature21056
7. Feng, W., Ju, L., Wang, L., Song, K., Ge, Z.: Towards novel class discovery: A study in novel skin lesions clustering. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J. et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 24–33. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-43987-2_3
8. Ferber, D., El Nahhas, O.S., Wölflein, G., Wiest, I.C., Clusmann, J. et al.: Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. Nature cancer **6**(8), 1337–1349 (2025)
9. Haenssle, H.A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T. et al.: Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology **29**(8), 1836–1842 (Aug 2018). https://doi.org/10.1093/annonc/mdy166
10. Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I. et al.: Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature Medicine **30**(9), 2613–2622 (Sep 2024). https://doi.org/10.1038/s41591-024-03097-1
11. Haggenmüller, S., Maron, R.C., Hekler, A., Krieghoff-Henning, E., Utikal, J.S. et al.: Patients' and dermatologists' preferences in artificial intelligence-driven skin cancer diagnostics: A prospective multicentric survey study. Journal of the American Academy of Dermatology **91**(2), 366–370 (Aug 2024). https://doi.org/10.1016/j.jaad.2024.04.033

12. Han, S.S.: SNU dataset + Quiz (3 2019). https://doi.org/10.6084/m9.figshare.6454973.v12
13. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C. et al.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding (2025), https://arxiv.org/abs/2510.08668
14. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets. IEEE Journal of Biomedical and Health Informatics **23**(2), 538–546 (Mar 2019). https://doi.org/10.1109/JBHI.2018.2824327
15. Kim, C., Gadgil, S.U., DeGrave, A.J., Omiye, J.A., Cai, Z.R. et al.: Transparent medical image AI via an image–text foundation model grounded in medical literature. Nature Medicine **30**(4), 1154–1165 (Apr 2024). https://doi.org/10.1038/s41591-024-02887-x
16. Kim, Y., Park, C., Jeong, H., Chan, Y.S., Xu, X. et al.: MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making (Oct 2024). https://doi.org/10.48550/arXiv.2404.15155
17. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H. et al.: LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day (Jun 2023). https://doi.org/10.48550/arXiv.2306.00890
18. Liopyris, K., Gregoriou, S., Dias, J., Stratigos, A.J.: Artificial Intelligence in Dermatology: Challenges and Perspectives. Dermatology and Therapy **12**(12), 2637–2651 (Oct 2022). https://doi.org/10.1007/s13555-022-00833-8
19. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X. et al.: A Survey on Hallucination in Large Vision-Language Models (May 2024). https://doi.org/10.48550/arXiv.2402.00253
20. Lyu, X., Liang, Y., Chen, W., Ding, M., Yang, J. et al.: WSI-Agents: A Collaborative Multi-Agent System for Multi-Modal Whole Slide Image Analysis. In: Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15964, pp. 680–690. Springer Nature Switzerland (September 2025)
21. OpenAI: Introducing GPT-5.2. https://openai.com/index/introducing-gpt-5-2/ (Dec 2025)
22. OpenAI, Hurst, A., Lerer, A., Goucher, A.P., Perelman, A. et al.: GPT-4o System Card (Oct 2024). https://doi.org/10.48550/arXiv.2410.21276
23. Pillai, J., Li, B.: Generative artificial intelligence in dermatology: Recommendations for future studies evaluating the clinical knowledge of models. Skin Research and Technology **30**(7), e13854 (Jul 2024). https://doi.org/10.1111/srt.13854
24. Ru, J., Yan, S., Yin, Y., Zou, Y., Ge, Z.: DermoGPT: Open Weights and Open Data for Morphology-Grounded Dermatological Reasoning MLLMs (Jan 2026). https://doi.org/10.48550/arXiv.2601.01868
25. Shen, Y., Sun, L., Xu, Y., Liu, W., Zhang, S. et al.: SkinCaRe: A Multimodal Dermatology Dataset Annotated with Medical Caption and Chain-of-Thought Reasoning (Nov 2025). https://doi.org/10.48550/arXiv.2405.18004
26. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**(1), 180161 (Aug 2018). https://doi.org/10.1038/sdata.2018.161
27. Wang, Z., Wu, J., Cai, L., Low, C.H., Yang, X. et al.: MedAgent-Pro: Towards Evidence-based Multi-modal Medical Diagnosis via Reasoning Agentic Workflow (Jul 2025). https://doi.org/10.48550/arXiv.2503.18968

28. Yan, S., Hu, M., Jiang, Y., Li, X., Fei, H. et al.: Derm1M: A Million-scale Vision-Language Dataset Aligned with Clinical Ontology Knowledge for Dermatology (Apr 2025). https://doi.org/10.48550/arXiv.2503.14911
29. Yan, S., Li, X., Hu, M., Jiang, Y., Yu, Z. et al.: MAKE: Multi-Aspect Knowledge-Enhanced Vision-Language Pretraining for Zero-shot Dermatological Assessment. In: Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15964, pp. 369–379. Springer Nature Switzerland (September 2025)
30. Yan, S., Yu, Z., Primiero, C., Vico-Alonso, C., Wang, Z. et al.: A multimodal vision foundation model for clinical dermatology. Nature Medicine **31**(8), 2691–2702 (Aug 2025). https://doi.org/10.1038/s41591-025-03747-y
31. Zeng, W., Sun, Y., Ma, C., Tan, W., Yan, B.: MM-Skin: Enhancing Dermatology Vision-Language Model with an Image-Text Dataset Derived from Textbooks. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3769–3778. MM '25, Association for Computing Machinery, New York, NY, USA (Oct 2025). https://doi.org/10.1145/3746027.3755187
32. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H. et al.: Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176 (2025)
33. Zhao, W., Wu, C., Fan, Y., Zhang, X., Qiu, P. et al.: An Agentic System for Rare Disease Diagnosis with Traceable Reasoning (Aug 2025). https://doi.org/10.48550/arXiv.2506.20430