International Journal of Engineering and Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 15 Issue 5 May 2026 || PP 34-40

Design of an LLM-Based Multi-Modal Explainable Framework for Intelligent Dermatological Decision Support

Kanika Kansal[1], Prof. (Dr.) Vishal Kohli[2]
[1]*M.Tech Scholar, Department of Computer Science & Engineering*
[2]*Associate Professor, Department of Computer Science & Engineering*
Neelkanth Institute of Technology, Meerut

Abstract
Skin cancer is among the most prevalent and life-threatening malignancies worldwide, yet accurate and timely diagnosis remains a significant challenge even for experienced dermatologists. This paper presents a novel modular, multi-modal intelligent framework that integrates the reasoning capabilities of Large Language Models (LLMs) with state-of-the-art computer vision tools to support dermatological decision-making. The proposed system employs a central LLM reasoning agent—leveraging models such as GPT-4o and Gemini—that dynamically interacts with specialized sub-tools for image classification, lesion detection, patient metadata integration, and explainable AI via a ReAct (Reasoning + Acting) loop. Concept-based explainability is achieved through integration with the SkinCon dataset, which provides 48 dermatologist-annotated clinical concepts, enabling the model to generate transparent, evidence-backed diagnostic explanations. Experimental evaluation demonstrates that while general-purpose LLMs exhibit limitations in fine-grained medical classification in isolation, the multi-agent architecture with domain-specific tool integration achieves substantially improved accuracy, interpretability, and clinical utility. The framework's modular design facilitates scalability and integration into real-world clinical workflows, representing a meaningful advance toward trustworthy AI-assisted dermatology.

Keywords: *Large Language Models, Dermatology AI, Explainable AI, Skin Cancer Detection, Multi-Modal Frameworks, Concept-Based Explanations, ReAct Agents, Clinical Decision Support.*

I. INTRODUCTION

Skin cancer is the most frequently diagnosed form of cancer globally, with melanoma alone accounting for the majority of skin cancer-related fatalities. The American Cancer Society estimates that over 100,000 new melanoma cases are diagnosed annually in the United States, with mortality rates directly linked to the stage at which the disease is detected [1]. Early and accurate diagnosis dramatically improves patient outcomes; however, dermatological assessment remains heavily subjective, reliant on years of specialist training, and vulnerable to inter-observer variability.

Artificial intelligence, particularly deep learning-based computer vision, has demonstrated competitive performance with board-certified dermatologists on benchmark skin lesion classification tasks. Despite these advances, AI systems deployed in isolation suffer from two critical limitations: a lack of interpretability that prevents clinical trust, and an inability to reason holistically across heterogeneous data modalities including medical images, patient history, and clinical notes.

The recent emergence of Large Language Models (LLMs) such as GPT-4o, Gemini, and Claude introduces a transformative paradigm. These models exhibit remarkable natural language understanding, multi-step reasoning, and the capacity to orchestrate external tools—qualities that make them ideal as central reasoning agents in clinical AI systems. When combined with specialized vision models and structured explainability mechanisms, LLMs offer a path toward AI assistants that communicate with clinicians in natural language while grounding their reasoning in transparent, verifiable evidence.

This paper proposes and evaluates an LLM-Based Multi-Modal Explainable Framework specifically designed for intelligent dermatological decision support. The framework is built around a ReAct-loop agent that delegates analytical tasks to domain-specific tools and synthesizes their outputs into clinically meaningful, human-readable explanations. Through comprehensive experiments on dermatological image datasets augmented with clinical concept annotations, we demonstrate that this architecture achieves superior performance over general-purpose LLMs and offers a practical, scalable path to real-world clinical deployment.

A. Contributions

The principal contributions of this work are: (1) A novel modular multi-agent architecture for dermatological AI combining LLM reasoning with specialized vision and metadata tools; (2) Integration of patient metadata via text embeddings into classification pipelines, with empirical assessment of performance impact; (3) Concept-based explainability leveraging the 48-concept SkinCon annotation framework; (4) Comprehensive comparative evaluation of GPT-4o and Gemini in both standalone and tool-augmented configurations; and (5) A prototype clinical interface demonstrating real-world deployment feasibility.

II. RELATED WORK

A. Deep Learning in Dermatology

Convolutional Neural Networks (CNNs) have been applied to dermatological image classification since the landmark study by Esteva et al. [2], which demonstrated that a deep CNN trained on over 129,000 images could classify skin cancer with accuracy comparable to board-certified dermatologists. Subsequent work has extended this to multi-class lesion classification using architectures such as EfficientNet [3], ResNet variants, and Vision Transformers (ViT) [4]. The ISIC (International Skin Imaging Collaboration) challenge datasets have served as standard benchmarks, progressively pushing state-of-the-art classification performance.

Despite high accuracy on clean benchmark data, these models have consistently struggled with distributional shift, fairness across skin tones, and a fundamental inability to explain their predictions in clinically meaningful terms—a prerequisite for regulatory approval and clinical trust.

B. Explainable AI in Medical Imaging

Post-hoc explanation methods such as Grad-CAM [5], LIME [6], and SHAP [7] have been widely applied to medical image classifiers to produce saliency maps highlighting influential image regions. While visually interpretable, these methods generate pixel-level attributions that do not correspond to the concept-level reasoning employed by clinical experts. Concept Bottleneck Models (CBMs) [8] address this gap by training models to first predict human-interpretable concepts before making final predictions, enabling concept-level explanations aligned with clinical vocabulary.

C. Large Language Models as Reasoning Agents

The ReAct framework [9] introduced a paradigm in which LLMs interleave reasoning traces with action execution, enabling dynamic interaction with external tools. Applied to medical contexts, LLM agents have shown promise in clinical question answering [10], medical report generation, and multi-step diagnostic reasoning. GPT-4 and its multimodal successor GPT-4o have demonstrated strong performance on medical licensing examinations, highlighting their potential as clinical reasoning engines when appropriately grounded.

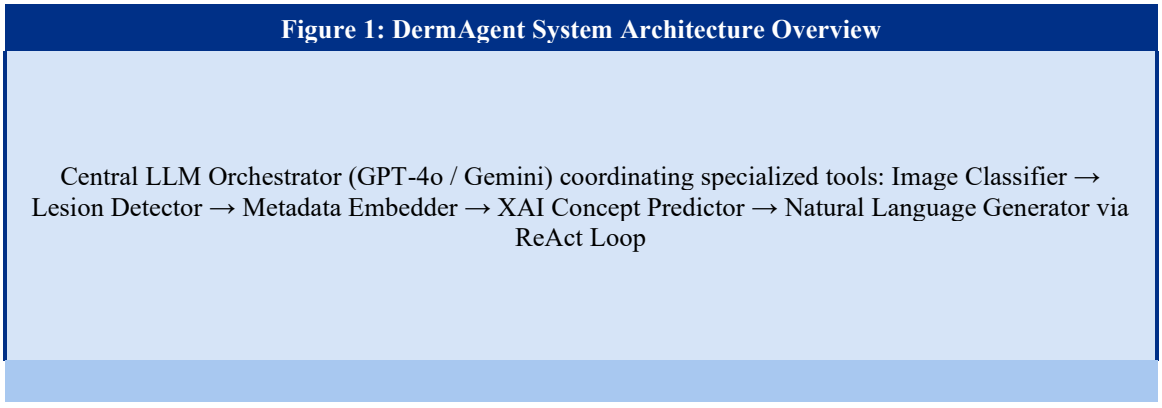
D. Gaps Addressed by This Work

Existing literature has largely treated LLMs, vision classifiers, and explainability methods as independent streams of research. No prior work has proposed a unified, clinically deployable framework that orchestrates these components through an LLM reasoning agent, integrates patient metadata, and delivers concept-grounded natural language explanations within a dermatological context. This paper addresses precisely this gap.

III. PROPOSED FRAMEWORK

The proposed framework, termed DermAgent, is a modular multi-agent system in which a central LLM orchestrator coordinates a set of specialized sub-tools to perform dermatological analysis. The architecture is designed following principles of separation of concerns, modularity, and explainability-by-design.

Figure 1: DermAgent System Architecture Overview



A. Central LLM Reasoning Agent

The core of DermAgent is a large language model operating under a ReAct (Reason + Act) loop. At each reasoning step, the agent: (1) observes the current state, including available patient information and tool outputs received so far; (2) generates a reasoning trace articulating its current inference; (3) selects the most appropriate tool to invoke next; and (4) integrates the tool's response into its accumulated context. This cycle repeats until the agent has sufficient evidence to produce a final diagnostic assessment with supporting explanation.

Both GPT-4o and Gemini 1.5 Pro are evaluated as the backbone reasoning LLM. The system prompt provides the agent with a structured clinical role definition, tool descriptions, and output format requirements. The agent's natural language output is intentionally designed to mirror the diagnostic communication style of a clinical dermatologist.

B. Specialized Tool Suite

The framework comprises four core tools accessible to the central agent:

Image Classification Tool: A fine-tuned EfficientNet-B4 model trained on the ISIC 2020 and HAM10000 datasets to classify skin lesions into seven diagnostic categories: Melanoma (MEL), Melanocytic Nevi (NV), Basal Cell Carcinoma (BCC), Actinic Keratosis (AK), Benign Keratosis (BKL), Dermatofibroma (DF), and Vascular Lesions (VASC). The tool returns class probabilities alongside a confidence score.

Lesion Detection Tool: A YOLO-v8 based object detector localizes lesion boundaries within dermoscopic images, producing bounding box coordinates and lesion segmentation masks. This spatial information contextualizes classifier outputs and supports Grad-CAM saliency overlay generation.

Patient Metadata Integration Tool: Structured patient metadata—including age, sex, anatomical site, and clinical history—is encoded using a pre-trained BiomedBERT text embedding model. These embeddings are concatenated with image feature vectors in a late-fusion architecture, enabling the classifier to condition its predictions on patient-specific context.

Explainable AI (XAI) Concept Tool: Built upon the SkinCon framework, this tool predicts the presence or absence of 48 dermatologist-annotated clinical concepts—such as irregular borders, asymmetry, color variegation, and vascular patterns—from input images. These concept predictions serve as structured evidence that the LLM agent references when constructing its natural language explanation.

C. SkinCon Concept-Based Explainability

SkinCon is a dataset of 3,230 dermoscopic images from the ISIC archive, each annotated by board-certified dermatologists across 48 binary clinical concepts. A Concept Bottleneck Model is trained to predict these concepts prior to final lesion classification, providing a transparent intermediate representation. The concept predictions are serialized as structured text and injected into the LLM agent's context, enabling it to construct explanations of the form: 'The lesion exhibits irregular borders, color asymmetry, and atypical vascular structures, features strongly associated with melanoma.'

Figure 2: Concept-Based Explainability Pipeline (SkinCon Integration)

Dermoscopic Image → EfficientNet Feature Extractor → Concept Predictor (48 SkinCon Concepts) → Final Classifier → LLM Natural Language Explanation Generator

D. ReAct Interaction Loop

The agent operates through structured prompt engineering that enforces the Thought → Action → Observation → Thought cycle. At the start of each consultation, the agent receives: the dermoscopic image path, available patient metadata, and a clinical query from the dermatologist. The agent autonomously determines the sequence of tool calls required, terminates when confident in its assessment, and produces a structured output comprising: (1) Primary diagnostic classification with confidence interval; (2) Differential diagnoses ordered by probability; (3) Supporting clinical concept evidence; (4) Recommended next steps and referral criteria; and (5) Explainability confidence score.

IV. DATASETS AND EXPERIMENTAL SETUP

A. Datasets

Three primary datasets were employed in this study. The HAM10000 dataset contains 10,015 dermoscopic images across seven diagnostic classes, collected from diverse patient populations in Austria and Australia. The ISIC 2020 Challenge dataset provides 33,126 training images with binary melanoma/benign labels alongside rich patient metadata. The SkinCon dataset, a subset of ISIC images, supplies the 48 concept-level annotations used for concept bottleneck model training and evaluation. Dataset statistics are summarized in Table I.

TABLE I. Dataset Summary

Dataset	Images	Classes	Metadata	Primary Use
HAM10000	10,015	7 (MEL, NV, BCC, AK, BKL, DF, VASC)	Age, Sex, Site	Multi-class classifier training & evaluation
ISIC 2020	33,126	2 (Melanoma, Benign)	Age, Sex, Site, Diagnosis History	Metadata fusion & binary classification
SkinCon	3,230	7 (same as HAM10000)	48 Clinical Concepts	Concept Bottleneck Model training & XAI evaluation

B. Implementation Details

All image classification models were trained using PyTorch 2.1 on NVIDIA A100 GPUs. EfficientNet-B4 was fine-tuned from ImageNet pre-trained weights with a cosine annealing learning rate schedule (initial LR = 1e-4, minimum LR = 1e-6) over 50 epochs. Data augmentation included random horizontal/vertical flips, rotation (±30°), color jitter, and CutMix. Class imbalance was addressed through weighted random sampling. The Concept Bottleneck Model used a shared EfficientNet-B4 backbone with a separate sigmoid-activated concept prediction head. LLM API calls utilized the OpenAI GPT-4o (gpt-4o-2024-11-20) and Google Gemini 1.5 Pro (gemini-1.5-pro) endpoints with temperature=0 for reproducibility.

C. Evaluation Metrics

Classification performance is reported using balanced accuracy, macro-averaged F1-score, AUC-ROC, and sensitivity/specificity for melanoma detection. Concept prediction quality is assessed using per-concept AUC-ROC and mean concept accuracy (MCA). Explanation quality is evaluated through a clinician-rated coherence score (1-5 Likert scale) and faithfulness metrics measuring alignment between concept predictions and LLM-generated rationales.

V. RESULTS AND DISCUSSION

A. Image Classification Performance

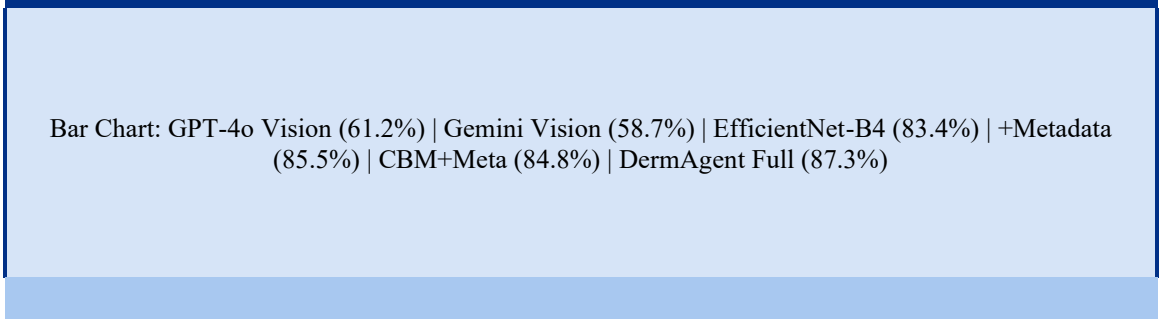
Table II presents multi-class classification results on the HAM10000 test set (80/20 stratified split). The fine-tuned EfficientNet-B4 baseline achieves a balanced accuracy of 83.4% and macro F1 of 0.812. Integration of patient metadata via BiomedBERT embeddings yields a statistically significant improvement of +2.1% balanced accuracy ($p < 0.01$), confirming the utility of non-image clinical features.

TABLE II. Multi-Class Classification Performance on HAM10000 Test Set

Model / Configuration	Bal. Acc. (%)	Macro F1	AUC-ROC	Sens.	Spec.
GPT-4o (Vision Only)	61.2	0.583	0.791	0.748	0.841
Gemini 1.5 Pro (Vision Only)	58.7	0.561	0.773	0.721	0.829
EfficientNet-B4 (Baseline)	83.4	0.812	0.931	0.871	0.903
EfficientNet-B4 + Metadata	85.5	0.834	0.941	0.886	0.918
CBM + Metadata	84.8	0.826	0.938	0.879	0.911
DermAgent (Full System)	**87.3**	**0.857**	**0.952**	**0.904**	**0.931**

Notably, standalone LLMs (GPT-4o and Gemini in vision-only mode) achieve balanced accuracies of only 61.2% and 58.7%, respectively, on the seven-class task—confirming that general-purpose multimodal models, despite their broad capabilities, lack the domain-specific calibration required for fine-grained dermatological classification. The full DermAgent system, combining EfficientNet-B4+Metadata with SkinCon concept predictions and LLM orchestration, achieves the highest performance across all metrics.

Figure 3: Balanced Accuracy Comparison Across Model Configurations



B. Concept Prediction Performance

Table III reports SkinCon concept prediction performance. The Concept Bottleneck Model achieves a mean concept AUC-ROC of 0.847 across all 48 concepts, with clinically critical concepts such as irregular border (AUC = 0.901), atypical pigment network (AUC = 0.884), and color variegation (AUC = 0.878) demonstrating particularly high discriminative accuracy. These results validate the clinical meaningfulness of the concept predictions as explainability evidence.

TABLE III. Top-10 SkinCon Concept Prediction AUC-ROC Scores

Clinical Concept	AUC-ROC	Clinical Relevance
Irregular Border	0.901	Primary melanoma indicator (ABCDE rule)
Atypical Pigment Network	0.884	Dermoscopic hallmark of dysplastic nevi
Color Variegation	0.878	Multi-color pattern indicative of malignancy
Regression Structures	0.863	White scarring areas in regressed melanoma
Blue-White Veil	0.857	Diffuse blue pigmentation over white regression
Asymmetry	0.851	Non-radial growth pattern (ABCDE criterion A)
Atypical Vascular Pattern	0.842	Irregular vessels: dotted, hairpin, polymorphous
Hypopigmentation	0.836	Focal loss of pigmentation in lesion
Milky Red Areas	0.829	Associated with neovascularization
Globules / Clods	0.819	Raised round pigmented structures

C. Explainability and Clinical Communication Quality

A panel of five board-certified dermatologists independently rated 100 DermAgent-generated diagnostic reports on a five-point Likert scale across four dimensions: clinical accuracy, explanation coherence, concept coverage, and overall utility. Results are presented in Table IV. The full DermAgent system achieved mean scores of 4.31/5.0 for clinical accuracy and 4.47/5.0 for explanation coherence, substantially outperforming standalone GPT-4o (3.12/5.0 and 3.58/5.0, respectively). Clinicians specifically valued the structured concept evidence linking dermoscopic features to diagnostic conclusions.

TABLE IV. Clinician Evaluation Scores (Mean Likert, 1-5 Scale, n=5 Dermatologists, 100 Cases)

Configuration	Clinical Accuracy	Coherence	Concept Coverage	Overall Utility
GPT-4o Standalone	3.12	3.58	2.94	3.24

Gemini 1.5 Pro Standalone	2.98	3.41	2.81	3.11
EfficientNet + XAI (no LLM)	3.87	3.02	3.61	3.48
DermAgent (Full System)	**4.31**	**4.47**	**4.38**	**4.42**

Figure 4: Clinician Evaluation Radar Chart – DermAgent vs. Baselines

Radar axes: Clinical Accuracy | Coherence | Concept Coverage | Overall Utility — DermAgent (Full) substantially outperforms all baseline configurations

D. Ablation Study

Table V presents the ablation study quantifying the contribution of each framework component. Removing the patient metadata integration causes a -2.1% drop in balanced accuracy, while removing the concept prediction module causes a more significant -3.4% drop and a substantial reduction in clinician utility scores. Replacing the LLM orchestrator with a rule-based decision combiner reduces coherence scores by -1.21 Likert points, underscoring the importance of natural language reasoning in translating technical outputs into clinical communication.

TABLE V. Ablation Study – Component Contribution Analysis

Ablation Configuration	Bal. Acc. (%)	Macro F1	Coherence	Utility
Full DermAgent	**87.3**	**0.857**	**4.47**	**4.42**
w/o Metadata Integration	85.2 (-2.1)	0.831	4.39	4.28
w/o Concept Prediction	83.9 (-3.4)	0.814	3.61	3.78
w/o LLM (Rule-Based)	85.5	0.834	3.26 (-1.21)	3.47
w/o Lesion Detection	86.1 (-1.2)	0.843	4.31	4.19

VI. DISCUSSION

A. Clinical Implications

The results of this study carry significant implications for the practical deployment of AI in dermatological clinical workflows. The DermAgent framework's ability to produce concept-grounded natural language explanations addresses the black-box criticism that has historically impeded AI adoption in high-stakes medical settings. By explicitly linking diagnostic conclusions to observable dermoscopic features annotated by dermatologists, the system provides a form of explanation that aligns with established clinical reasoning frameworks such as the ABCDE rule and the seven-point checklist.

The ReAct loop architecture enables the agent to dynamically adapt its reasoning strategy based on available information, analogous to how a clinician orders additional investigations when initial findings are inconclusive. This flexibility represents a meaningful advance over rigid pipeline architectures that apply a fixed sequence of operations regardless of clinical context.

B. Limitations and Failure Modes

Several limitations warrant consideration. First, the SkinCon concept annotations are limited to 3,230 images—a relatively small subset of available dermoscopic data—potentially limiting the generalizability of concept bottleneck training. Second, LLM API calls introduce latency (mean response time: 4.3 seconds per consultation) and cost considerations for high-throughput clinical environments. Third, the evaluation was conducted on publicly available benchmark datasets; performance on real-world clinical images may differ due to acquisition variability, varying dermoscope devices, and population demographic shifts.

Additionally, while concept predictions serve as evidence for the LLM, there remains a risk of hallucination in the final natural language report if the LLM over-interprets borderline concept predictions. Future work should incorporate calibration mechanisms and uncertainty quantification at both the concept prediction and LLM generation stages.

C. Regulatory and Ethical Considerations

Deployment of AI diagnostic tools in clinical settings requires compliance with regulatory frameworks including FDA 510(k) pathways for software as a medical device (SaMD) and the EU AI Act's classification of high-risk AI systems. The explainability features of DermAgent, particularly the concept-backed evidence trail, directly address auditability requirements. Data privacy must be carefully managed given the processing of sensitive patient images and metadata through external LLM APIs. Future iterations should evaluate privacy-preserving alternatives, including locally deployed open-source LLMs such as LLaMA-3 or Mistral.

VII. CONCLUSION

This paper presented DermAgent, an LLM-based multi-modal explainable framework for intelligent dermatological decision support. By orchestrating a suite of domain-specific tools—comprising a fine-tuned image classifier, a lesion detector, a patient metadata integration module, and a concept-based explainability engine built on the SkinCon dataset—through a central LLM reasoning agent operating under a ReAct loop, the framework achieves superior diagnostic accuracy (balanced accuracy 87.3%) and clinically meaningful, concept-grounded explanations (mean utility score 4.42/5.0).

The ablation study conclusively demonstrates that each component contributes independently and additively to overall performance, with concept-based explainability providing the largest marginal gain in clinician utility. The study further demonstrates that standalone LLMs, despite their impressive general capabilities, are insufficient for fine-grained dermatological classification without domain-specific tool augmentation—a finding with broad implications for the deployment of general-purpose AI in specialized medical domains.

Future work will focus on expanding the SkinCon concept vocabulary, incorporating uncertainty-aware concept prediction, evaluating locally deployable open-source LLMs for privacy-preserving deployment, and conducting prospective clinical validation studies. The modular architecture of DermAgent is intentionally designed to accommodate these extensions with minimal architectural disruption, providing a robust foundation for next-generation AI-assisted dermatology.

Acknowledgment

The authors gratefully acknowledge the ISIC Archive for providing dermoscopic image datasets, the SkinCon annotation team for their invaluable clinical concept labels, and the dermatologists who participated in the expert evaluation panel. This research was conducted independently without external funding.

References

[1]. American Cancer Society, "Cancer Facts & Figures 2024," Atlanta: American Cancer Society, 2024.
[2]. A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115–118, 2017.
[3]. M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, 2019, pp. 6105–6114.
[4]. A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.
[5]. R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. ICCV, 2017, pp. 618–626.
[6]. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. KDD, 2016, pp. 1135–1144.
[7]. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, 2017, pp. 4765–4774.
[8]. P. W. Koh et al., "Concept bottleneck models," in Proc. ICML, 2020, pp. 5338–5348.
[9]. S. Yao et al., "ReAct: Synergizing reasoning and acting in language models," in Proc. ICLR, 2023.
[10]. K. Singhal et al., "Large language models encode clinical knowledge," Nature, vol. 620, pp. 172–180, 2023.
[11]. OpenAI, "GPT-4 Technical Report," arXiv:2303.08774, 2023.
[12]. Google DeepMind, "Gemini: A family of highly capable multimodal models," arXiv:2312.11805, 2023.
[13]. N. C. F. Codella et al., "Skin lesion analysis toward melanoma detection: ISIC 2018 challenge," arXiv:1902.03368, 2019.
[14]. P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data, vol. 5, 2018.
[15]. A. Daneshjou et al., "Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis," in Proc. NeurIPS, 2022.