SkinBench : A Multimodal LLM Benchmark for Skin Disease Diagnosis

Submitted By,

SWARNA AKTER
ID : 221-35-919

Supervised By,
Md. Shohel Arman
Assistant Professor
Department of Software Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

APPROVAL

This thesis titled on "**SkinBench : A Multimodal LLM Benchmark for Skin Disease Diagnosis**", submitted by **Swarna Akter (ID: 221-35-919)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. Imran Mahmud
Professor & Head
Department of Software Engineering
Faculty of Science and Information Technology Daffodil International University

Internal Examiner 1

Afsana Begum
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Md. Shohel Arman
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Nadira Islam
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Md Manowarul Islam
Professor
Department of Computer Science and Engineering
Jagannath University, Bangladesh

Department of Software Engineering

Faculty of Science and Information Technology

Supervisor Approval Form

Fall 2025	B.Sc. In SWE	Campus: DSC

Student Name	Student ID
Swarna Akter	221-35-919

Project/Thesis Information	
Thesis Title	SkinBench : A Multimodal LLM Benchmark for Skin Disease Diagnosis
Type of work	Thesis
Supervisor information	
Supervisor Name	Md. Shohel Arman
Supervisor Initial	MSA
Completed Credit till now	133
How many credits in this semester	6
Amount (Due)	0
Supervisor Consent	Yes ☐ No ☐

Supervisor Signature

DAFFODIL INTERNATIONAL UNIVERSITY

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	Swarna Akter
Date of Birth	:	08-01-2002
Title	:	SkinBench : A Multimodal LLM Benchmark for Skin Disease Diagnosis
Academic Session	:	2022-2025

I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☐	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:

(Student's Signature)

221-35-919

Student ID
Date: 24-12-2025

(Supervisor's Signature)

Md. Shohel Arman

Name of Supervisor
Date:

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.

(Supervisor's Signature)

Full Name : Md. Shohel Arman

Position : Assistant Professor

Date : 24 December 2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

(Student's Signature)

Full Name : Swarna Akter

ID Number : 221-35-919

Date : 24 December 2025

SkinBench : A Multimodal LLM Benchmark for Skin Disease Diagnosis

Submitted By,

SWARNA AKTER

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering (Major in Data Science)

DAFFODIL INTERNATIONAL UNIVERSITY

DECEMBER 2025

ACKNOWLEDGEMENTS

All praise and thanks are due to Allah (SWT), the Most Gracious and the Most Merciful. His guidance, blessings, and protection have helped me reach this stage. Without His mercy, the completion of this thesis would not have been possible.

Now I would like to say my heartfelt thanks to my supervisor, Md. Shohel Arman. Assistant Professor, due to his constant support, valuable advice, and power. scholarly assistance during this study. his forbearance, his focus, and worth. feedback contributed to the fact that I could improve my work in each step. I am really thankful to the time and effort he spent in directing me. I owe my family much gratitude to their constant encouragement, understanding and. prayers. Their assistance has provided me with strength and confidence in the most adverse times. moments of this journey. I am also grateful to my friends because they encouraged me. positive words, which assisted me in being motivated during this work. Finally I would like to thank all those who helped, either directly or indirectly, the. successful accomplishment of this thesis. Their support and goodwill have meant a great deal to me.

DEDICATION

Dedicated to all individuals living with skin diseases, those who face daily discomfort, emotional pain, and social challenges with quiet strength. This work is also for those who lack the resources for proper diagnosis or treatment. May this research contribute, even in a small way, to easing their path.

ABSTRACT

There is an emerging trend to use Large Language Models (LLMs) in medical diagnosis (particularly analysis of complex data such as medical imaging and patient histories). Nevertheless, currently, models tend to be not as deep in reasoning and as well as clinically accurate when used in real-life scenarios of healthcare. To solve this, we introduce SkinBench, which is a new benchmark and assessment framework that is concentrated on diagnostics of skin diseases. It is based on a multi-agent system comprising of three specialized agents they are (1) DescribeLLM which describes the clinical scenario; (2) DoctorLLM which acts as a clinician asking questions and reasoning through a diagnosis and (3) EvalLLM which assesses the quality of the diagnostic outcome. SkinBench has a total of 500 cases of skin diseases, which consist of both images and written reasoning, as well as the model-generated dialogues, and responses- 10 percent are checked by the real doctors. The first screening step is to first determine the standalone diagnostic accuracy of seven representative LLMs on these 500 cases: GPT-5.1 has the highest accuracy of 98.8, followed by Mistral with 98.6, GPT-4o with 95.6, DeepSeek with 94.0, Qwen with 91.4, Llama 3.2 with 90.6, and GPT-3.5 Turbo with the lowest accuracy of 88.6. Such ranking reveals that there exists significant differences in performance among models prior to the implementation of multi-agent reasoning and it is therefore pertinent that a more organized and realistic benchmark is developed. We thus evaluate seven common skin diseases in Bangladesh, Scabies, Psoriasis, Monkeyopx and Chickenpox, Candidiasis and Tinea, Atopic Dermatitis and Seborrheic Dermatitis, Acnes and Impetigo, and assess both open-source and closed-source LLMs on three main dimensions: (1) the accuracy of the diagnostic results of skin diseases, (2) the consistency of reasoning among multi-turn conversational interactions, and (3) the quality and explainability of clinical results.

Keywords - Large Language Models, Dermatology, Benchmarking, Multimodal Reasoning, Medical AI, Diagnostic Evaluation, Multi-Agent Framework, Skin Disease.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

LLM = Large Language Model

QA = Question Answering

CNN = Convolutional Neural Network

LIST OF APPENDICES

Appendix A:

CHAPTER 1

INTRODUCTION

1.1 Background

Large Language Models have shown impressive and fast-growing opportunities in healthcare applications and shifted the nature of healthcare systems in dealing with medical diagnosis, clinical decision support, and patient care delivery [1, 2]. Recent transformative progress in the field of natural language processing and multimodal machine learning has allowed the use of LLMs to simultaneously process and intelligently combine diverse clinical information such as high-resolution medical imaging, detailed patient histories, detailed descriptions of symptoms, past diagnostic examinations, and clinical reasoning patterns clinicians follow during disease diagnosis [3, 4]. These technological developments have caused a significant amount of attention and extensive research spending on using LLMs to specialized medical areas, where clinical knowledge, diagnostic precision is vital to patient outcomes, and expert access is extremely limited, especially in developing nations such as Bangladesh. In Bangladesh in particular, there is a gross shortage of dermatological knowledge and unequal distribution among the healthcare network, and originates serious diagnostic obstacles especially beyond the larger urban centers. The geographical and occupational imbalance in distribution leads to a delay in diagnosis, avoidable and untreated misdiagnosis, and poorer treatment of millions of Bangladeshis with skin disorders. The country is experiencing huge disease burdens due to skin diseases which have been very mostly predominant in tropical and subtropical areas with tropical climates, overpopulated cities, and inadequate sanitation structures. Examples of common skin diseases include Scabies--a parasitic skin infection that is highly prevalent among the crowded group of populations with limited access to sanitation; Psoriasis- a chronic inflammatory skin condition, requiring specialized treatment; Monkeypox and Chickenpox- viral infections with high morbidity; Candidiasis and Tinea- fungal infections that are especially common among the hot and humid tropical climate of Bangladesh; Atopic Dermatitis and Seborrheic Dermatitis- chronic inflammatory dermatoses with All these conditions are causes of millions of Bangladeshi patients but are poorly and sometimes misdiagnosed because people very often do not have access to specialized dermatological knowledge, have little diagnostic equipment in primary healthcare centers, and can be geographically isolated in relation to the specialist.

Recent advances in multimodal Large Language Models have shown potential opportunities in the application of dermatological diagnosis. A complex multimodal system, referring to SKGPT-4, combining vision transformers and Llama-2-13b-chat language model architecture demonstrated convincing work on the 150 real-world clinical cases in which the diagnosis is independently confirmed by board-certified dermatologists, which provides evidence of concept proof with the use of LLM-based dermatological diagnosis. In-depth medical benchmarking systems and other systems such as CliBench and COGNET-MD have developed stringent and organized evaluation approaches to the evaluation of the performance of LLM in various medical activities such as disease diagnostics, generation of treatment recommendations and clinical reasoning abilities [8, 9].

Nevertheless, regardless of these innovations, existing Large Language Models are limited to significant and reported issues in use in clinical contexts [10]. These challenges involve: first, LLMs often have difficulty in keeping chains of logical reasoning consistent sequentially, over diagnostic conversations, which is an essential feature of realistic clinical practice with diagnosis developing as conversations proceed through repeated questioning, hypothesis testing, and information combination across many conversational events [9, 10]. Second, the available LLCMs are not optimized to specific domains, such as dermatological diagnostics, and thus their application to dermatological diseases is not the most optimal [3, 4]. Third, LLMs often make hallucinated diagnoses, which are artificial clinical data, associations between symptoms, or fabricated clinical facts, which can threaten patient safety unless applied in clinical practice under the supervision of a human expert. Fourth, the existing LLM assessment models were created mainly in the context of Western medicine and disease manifestations rather than disease epidemiology in Bangladesh. Importantly, an exhaustive assessment model does not exist to help the clinicians in Bangladesh and potentially in other parts of the world that deal with skin ailments common in Bangladesh or offer specifications to the clinicians who may apply the LLM-based diagnostic support systems in the Bangladesh healthcare clinical setting.

1.2 Problem Statement

The current state of general-purpose LLMs in healthcare is encouraging in terms of single-turn question answering, but frequently cannot maintain consistent and coherent reasoning on multi-turn dermatology conversation. In long discussions, models can prove inconsistent with their previous words, lose the symptoms or risk factors described before, and alter their diagnostic hypothesis with no sufficient reason, which does not contribute to the credibility of clinicians in their recommendations. These are particularly problematic in dermatology as in this case, minute aspects of timing, distribution, and the behavior of symptoms are important in making the right diagnosis.

Besides, the memory constraints of LLMs mean that vital context may be lost or misconstrued as the conversation becomes lengthy, particularly when patients present more than one complaint relevant to the conversation or when the follow-up inquiries take up many turns. This memory impairment causes the incomplete chain of reasoning and may cause the neglect of red-flag signs as rapidly progressing lesions, systemic manifestations, or history of malignancy, which predisposes the patient to unsafe or delayed diagnoses.

Existing models are further weak in having strong understanding of regional patterns of skin disease, environmental exposures and cultural practices affecting people such as Bangladesh. Since majority of the training and evaluation information is based on other areas, the models will misclassify ordinary local conditions, over-rank some uncommon Western illnesses, or give guidance that is not relevant to local health facilities and patient facts. The misfit increases the issue of fairness, applicability, and safety when such models are applied in the low- and middle-income contexts.

Furthermore, a special skin-disease benchmark, which can systematically compare various LLMs and multi-agent systems with tasks of realistic dermatology, is yet to be provided. In the absence of a unified collection of multimodal instances, dialogue structures, and

reasoning- sensitive measures, there is little reason to measure the success and failure of models, or trace the progression of successive generations of LLMs. This is why the results on complex dermatology cases are still in most cases not entirely consistent with one of the models providing plausible but incorrect diagnosis and other models providing superficially fluent but clinically unsound reason.

Combined with these restrictions, these factors drive the desire to have a dermatology-oriented, geographically sensitive benchmark and a corresponding multi-agent model that explicitly aims at the quality of reasoning, memory during extended conversations, geographical disease information, and stability over complicated, real-world skin problems.

1.3 Research Gap

Recent work on medical LLMs tends to evaluate models on broad, mixed-specialty benchmarks or general clinical QA, rather than on disease-specific settings where domain nuances matter strongly, such as dermatology. As a result, dermatology is rarely treated as an independent evaluation domain, and there is limited understanding of how LLMs handle lesion description, morphology, and pattern-based reasoning that are central to skin diagnosis. Existing datasets and leaderboards in medical AI focus largely on internal medicine, radiology, or general clinical questions, leaving a clear gap for systematically testing LLMs on cutaneous disorders across diverse case types and difficulty levels.

Furthermore, there is no dedicated, widely used benchmark that combines multimodal input (clinical photographs plus text) with multi-turn dialogue and explicit reasoning traces for skin diseases. Most available dermatology datasets were originally designed for image-only classification or for training CNN-based models, and they do not capture the conversational, question-and-answer nature of real dermatology consultations where history, risk factors, and evolution over time are crucial. This limits the ability of researchers to fairly compare LLM-based pipelines that integrate image descriptions, follow-up questions, and diagnostic explanations.

Another important gap is the lack of region-aware benchmarks that represent common skin conditions in specific populations, such as Bangladeshi patients. Many existing datasets are biased toward Western cohorts and high-resource hospital settings, which under-represent infections, pigmentary disorders, and environmental or occupational dermatoses that are more prevalent in South Asia. Without region-specific evaluation, it is impossible to know whether an LLM that performs well on global datasets can safely support clinicians in local contexts.

Finally, although multi-agent architectures have been explored in general clinical decision support, there is almost no work on dermatology-optimized agent roles and collaboration patterns. Prior systems typically use generic “doctor” and “reviewer” agents without tailoring them to tasks such as lesion description, differential diagnosis generation based on morphology, or checking for high-risk skin cancers. This leaves open questions about how to design, coordinate, and evaluate specialized agents that can jointly improve diagnostic reasoning and safety for skin disease cases.

1.4 Research Objectives

This thesis aims to create a primary benchmark and evaluation framework focused on dermatology for LLM-based skin disease diagnosis. It will revolve around realistic multi-turn. consultations rather than shot-type questions. The benchmark will comprise of both multimodal. input -clinical images and textual histories and clear reasoning traces. This way, models will be assessed on the interpretation of lesions, collection of the corresponding information and justification of the same. diagnostic choices. Among them is the objective to base the benchmark on disease patterns that are pertinent to. the region. The large percentage of the cases will be typical skin conditions in Bangladeshi. and presentation styles to solve the existing imbalance with Western datasets.

The work will also build a hybrid dataset that merges carefully selected open or synthetic cases with a controlled set of real clinical data. About 15 to 20% of the benchmark cases will come from anonymized hospital records and image archives, following institutional approval and strict de-identification. To better reflect realistic questioning behavior, around 10% of the benchmark dialogues will include questions and reasoning steps that practicing dermatologists wrote or reviewed, providing high-quality examples for assessing and guiding model behavior. This thesis also aims to create and evaluate a dermatology-focused multi-agent LLM framework built around three coordinated agents—DescribeLLM, DoctorLLM, and EvalLLM—using the new skin disease benchmark.

Another goal is to use the benchmark to systematically evaluate multiple closed-source and open-source LLMs. This evaluation will measure not only exact diagnostic accuracy but also clinically acceptable alternatives, reasoning quality, suitability of suggested investigations, and overall safety of recommendations. The aim is to identify which interaction patterns and model types work best for dermatological decision support and where specific improvements are still needed for safe deployment.

1.5 Research Contributions

This work aims to make LLM-based dermatology support safer and more reliable. It does this by introducing a benchmark focused on dermatology, a hybrid real-world dataset filled with expert-written dialogues, and a specialized evaluation framework for diagnosing skin diseases. First, it presents a benchmark designed to evaluate large language models specifically for diagnosing and reasoning about skin diseases, rather than for general medical questions. This standard incorporates multimodal cases in which one of the clinical skin images is at minimum. is accompanied by a structured text such as the chief complaint, medical history and examination. findings. It has a multi-turn dialogue template, too, which entails models to pose questions, as well. test their hypotheses, and support their eventual diagnoses. Second, the benchmark is created as a hybrid dataset, which is based on clinical practice. It is curated synthetic and textbook-style, and 15-20% of the cases. derived out of anonymized hospital data and dermatologic image bank. This inclusion introduces real types of lesions, comorbidity, and other documentation varieties into the assessment. and without violating patient privacy by maintaining anonymity. Approximately 10% of the conversation includes consequent questions and lines of thought that are either penned or attentively. verified by dermatologists on duty. These are professional contributions to how clinicians. ask questions, eliminate differentials, and find red-flag cases. They also serve as high-quality

references for assessing model performance.

Third, the thesis proposes a multi-agent framework designed to suit dermatology which involves three collaborating agents: DescribeLLM, DoctorLLM and EvalLLM. This framework structure the process of diagnosis into clinical summarization, multi-turn diagnostic reasoning, and systematic evaluation. In this type of structure, DescribeLLM transforms the raw image and text. information into an explicit clinical summary. DoctorLLM proceeds with the successive consultation, generating the differential diagnosis, proposed investigation, and treatment plans. EvalLLM then evaluates the whole interaction on such factors as the correctness of the diagnosis, coherence. of reasoning, suitability of tests and safety of suggestions. Lastly, the thesis utilizes the new benchmark and multi-agent framework to perform. robustness experimenting with a number of open-source and closed-source LLMs. It also contrasts the example of a single agent of the DescribeLLM-DoctorLLM-EvalLLM pipeline, with the same cases of dermatology. This comparison reveals the discovery of the benchmark. distinctions, which standard measures of accuracy fail to capture, give support to the fact that the multi-agent design, a dermatological oriented design improves the retention of context during multi-turn. exchanges and generates more rational and harmless thinking to diagnose skin diseases. This work therefore opens up the way to the specialized use of LLM in dermatology.

1.6 Motivation

The large language models (LLMs) have demonstrated a high level of understanding and generation skills. natural language in most spheres such as general knowledge and domains. Recently, The application of LLMs and healthcare along with medical decisions has become of interest. Models similar to GPT-4 and Claude and other open-source alternatives are also doing well on medical. examinations and clinical reasoning. Yet, even though they are gradually gaining presence in healthcare AI, the absence of comprehensive assessment of the abilities of LLM remains. Particularly in the diagnosis of dermatology. This is a medical specialty that is dependent on visual cues. and pattern recognition. It would be important to establish benchmarks to measure and compare performance of fairly. various models of straightforward, standardised tasks. The present-day medical AI standards, including PubMedQA, MMLU-Professional and MedQA, are predominantly general medical knowledge. by means of text-based question-answer methods. These standards evaluate models in terms of facts. recall and multiple choice reasoning but in many cases do not represent complexity of real world. clinical practice. Doctors are involved in multi-turn interaction, pose follow-up questions. up questions, and make adjustments to their diagnoses on new information about the patient. Dermatology poses a distinct challenge. It is diagnosed based on its recognition of visual patterns and conversion. observations into visual descriptions of clinical arrangements. It also involves the necessity of reasoning through. these descriptions in discourse. The available benchmarks fail to capture this process, rendering it challenging to comprehend the level of ability of LLMs to complete the particular mentally pertinent task. dermatological reasoning. This entails the transformation of clinical observations into diagnostic. questions and a logical consistency between provision of questions, provision of answers. and clinical reasoning. Moreover, even though dermatology can be considered one of the most widespread medical fields. there has been worldwide, with billions of people being afflicted with skin conditions annually. regrettably, little has been done in terms of specific LLM benchmarks used to diagnose skin diseases. Most current medical AI benchmarks treat dermatology as a minor part of general medicine. This results in few cases being considered and limited depth in evaluating dermatological reasoning, leads to

minimal cases to be taken into account and reduced profundity in the assessment of dermatological. reasoning. This underrepresentation fails to represent the importance of dermatology in primary. healthcare and possible ways AI could enhance the access to providing the correct diagnosis of skin diseases in. areas with fewer resources. In order to solve these weaknesses, this work constitutes a new standard particularly in the evaluation. Performance in dermatological reasoning by LLM. Rather than just ascertaining whether models can. One of the multiple-choice evaluations is the benchmark of how well one chooses the correct diagnosis on a list of multiple choices. LLMs are capable of engaging in multi-turn conversations about diagnostic processes in a real world. The process converts processes skin lesion images to objective clinical observations, with GPT-4o producing model-specific, and converts them to image form. diagnostic questions, and checks that the answers of the model are correct as well as. intuitively coherent with its presented logic. This two part validation right and consistency of reasoning - makes sure that the benchmark really assesses clinical reasoning ability. not a haphazard guess-work or a mutual contradiction.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The chapter represents an overall review of the existing literature in the field that is relevant to crucial areas: the best practices of Large Language Models used in medical diagnostics, current evaluation models and standards of evaluation, specialized systems, medical LLMs designed to be used to diagnose dermatology specially, systems facilitating multi-agent cooperation in LLM systems, and assessment ways of rigorously evaluating clinical reasoning in multi-turn conversational settings.

2.2 Multimodal and Tool-Augmented LLMs in Medicine

The recent research has expanded large language models into multimodal and tool-augmented. medical systems which can mix clinical text with structured information, images, and external. tools. DISC-MedLLM is an interface between the general-purpose LLMs and the medical world consultation through a combination of retrieval, domain-specific knowledge and structured reasoning templates, which demonstrate that careful grounding can make diagnostic significantly better. During pure black-box prompting is unstable [14]. This idea is further developed by MedAgent-Pro through a multi-modal framework that is evidence-based where planning agents liaise has image analysis, clinical text understanding and external tool calls, matching the reasoning workflow based on guidelines, as opposed to free generation [12]. As MMedAgent shows, a multimodal agent architecture is capable of learning when and how to appeal to expert medical equipment (e.g., radiography, laboratory analysis) and that such equipment-conscious orchestration demonstrates the best performance in various medical activities as compared to autonomous vision-language models and powerful general LLMs [13]. Healthcare-Agents makes the vision bigger by placing LLMs as the key drivers of health prediction and decision-making pipelines, pointing out potential paths of incorporating predictive conversational interfaces, EHR signals, and models, as well as highlighting unresolved problems with strength and security [21]. Combined, these papers demonstrate that multimodal, tool-augmented LLMs offer an excellent starting point, which can be targeted at making clinically oriented assistants. They too indicate the necessity of a better organized regulation of the reasoning steps and interactions with external components [12–14][21].

2.3 Multi-Agent Frameworks for Medical Diagnosis

Multi-agent architectures have been introduced as an interesting means of introducing structure upon complicated diagnostic work, where various LLM-based agents specialize, including planner, clinician, tool-caller and critic [12][14][15]. Overall surveys of multi-agent systems based on LLM outline patterns of collaboration, communication protocols as well as adaptable coordination mechanisms applicable to high stakes areas such as healthcare [16][18]. Surveys of agents using LLM in medicine and AI hospitals in the medical context

list typical design factors, such as task division and role specialization. Overall, these domains can be enhanced through the use of multi-agent collaboration, shared memory, and tool access, interpreting and performing better than single-agent baselines [17][18]. These concepts are depicted in a number of concrete medical multi-agent frameworks. MEDAIDE suggests an omni-medical assistant which breaks down user requests into sub-tasks managed by professional agents (e.g., pre-diagnosis, diagnosis, medication, post-care), orchestrated by the recognition of the intent and deliberate stage changes [11]. MedAgent-Pro and DISC-MedLLM have hierarchical planning and implementation, in which high-level planners develop diagnostic programs and sub-agent telephones or create descriptions that contain feedback loops that impose evidence-based thinking [12][14]. MDTeamGPT introduces a multi-agent multi-disciplinary team framework that is self-evolving consultation, which enables virtual specialists to discuss a case, update their opinion, and agree on a diagnosis and treatment plan in the long run [15]. Self-Evolving Multi-agent Simulations build on this concept by allowing the agent society and interaction patterns to be developed as a result of repeated clinical simulations, in order to be more realistic, adaptive collaboration dynamics [17]. All these works are a pointer in the fact that multi-agent structuring may reflect characteristics of actual clinical practice, such as expertise division and peer review, that single-agent model is not easily replicated [11–15][17].

2.4 Benchmarking and Evaluation of Medical AI Agents

Benchmarks to assess the performance of LLM-based agents have been suggested to be performed by simulation, not in question-answering but in realistic and multi-turn circumstances. Agent-Hospital simulates a virtual hospital where evolvable medical agents are interacting with simulated patients, where the diagnostic accuracy, treatment decisions, and can be assessed. The unfolding of communication strategies cases is multiturned [19]. AI-Hospital builds on this notion to compare a broad range of LLMs in a hospital-like scenario, with the emphasis on it regarding the quality of consultations, safety, and compliance with the clinical norms, and disclosing big diversification in the performance of models even when they are similar in the traditional NLP benchmarks [20]. In simulated clinical, AgentClinic proposes a multimodal agent benchmark of AI environments, having explicit roles, including, doctor, patient, measurement, and moderator agents, and a simulation of the real-world injected patient and physician bias. The benchmark measures the diagnostic accuracy, along with monitoring the agent behavior with respect to noisy information, cognitive bias, and uncertainty, and thus bringing appraisal nearer to realistic clinical practice [28]. Survey of a larger scope on multi-agent AI hospitals based on LLM combines these platforms, contrasting their architectures, task coverage and metrics, and stresses, interactive, role-based simulations are pivotal to the comprehension of how medical agents do not act on through a single-turn QA tests [18]. Nevertheless, there are current simulations are relatively specialty-agnostic and they do not provide a dermatology-specific case mixes or measures of reasoning, which vacates to the skin-disease-oriented measures [11][18].

2.5 Large Language Models in Medical Diagnosis: Current Capabilities and Performance

Large Language Models show high and steadily reported performance on medical knowledge tasks, clinical reasoning problems [1, 21]. GPT-4 achieves significant passing of medical licensing examinations such as the United States Medical Licensing Examination, Spanish medical resident exams and other standardized exams medical assessments [1, 20]. Claude

3.5 has good medical reasoning ability with proper clinical background and logic. Open-source has been shown to be effective with competitive alternatives such as Llama 3.2 with different parameter counts, hope of medical uses, and especially with proper reproduction and optimization in the case of certain medical fields. Understanding of model selection has shown conclusively that the selection of models is affected by the research problem itself, has a considerable effect on the performance of the clinic, and domain-specific aspects is becoming more significant as the level of medical specialization and task complexity rises [1, 21]. Abilities of various model families differ considerably. Closed-source proprietary models tend to perform better on medical tasks in terms of absolute performance because of high superiority train methodologies, bigger training data and reasoning maximization tasks [20, 23, 21]. Nevertheless, open-source options gain more competitive choices available performance that is highly beneficial in terms of interpretability, cost-effectiveness and deployment flexibility that can be deployed to resource-constrained environments [24, 32].

2.6 Surveys and Collaboration Mechanisms

Several surveys give a more advanced picture of the agents based on LLM in medicine and mechanisms of multi-agent collaboration. Summaries of multi-agent systems based on LLC characterize coordination plans (e.g. debate, voting, role-based workflow), architectural patterns, and common application areas, sketching the design dimensions, e.g., sharing of memory, centralization and access to tools [16][18]. A dedicated survey on the use of LMMS-based agents in medicine examines clinical use scenarios, safety issues, regulatory issues, and standard agent functions (planner, reasoner, retriever, explainer), emphasizing the necessity of open thinking and human control in the high-stakes settings [17]. The other survey of AI hospitals and medical LLM agents systematically compares existing simulated hospital platforms, clinical planning mechanisms and assessment measures, discovering the absence of specialty-specific and region-sensitive benchmarks as one of its limitations [18]. In general, these surveys all suggest that: (i) medical agent systems must be designed around clear workflows that are connected to clinical practice; (ii) assessment should take into account the quality of reasoning, safety and cooperation, rather than the correctness of answers; and (iii) domain-specific tests must be available to reveal modes of failure not detected by generic tests [16–18]. This encourages the creation of region-sensitive and dermatological benchmarks and frameworks of agent as in the one suggested in this thesis.

2.7 Learning to be a Doctor and Adaptive Workflows

In addition to fixed architectures, there are also studies of adaptive medical agent systems that are capable of get to know how to work better with time. Automated search is presented in Learning to Be a Doctor using performance feedback on the roles of over agents, communication patterns, and the use of each tool optimize the architecture and rules of interaction [25]. Multi-agent Simulations Self-Evolving equally permit the structure and action of agents to evolve with their involvement in clinical situations reused, towards more realistic and resilient interaction patterns [17]. Healthcare-Agents is a perspective that views LLMs as agents of coordination of predictions, tools, and out-of-control models throughout the entire health-prediction and decision-making pipeline, indicating that adaptive coordination will become essential to the long-term implementation in inhomogenous clinical settings [21]. These guidelines lead to the future systems that do not only adhere to pre-defined workflows but they also know how to construct their own thinking and team work, a concept that SkinBench uses in designing interactions between dermatology-specific agents.

2.8 Gaps for Dermatology and Skin-Disease Benchmarks

In these works, dermatology is typically mentioned as one of the numerous areas of test, and it has no special benchmark aiming at skin-disease diagnosis with multimodal, turn-taking reasoning and region specific case distributions [11-13]. Current agent and hospital simulations are infrequently given detailed cases of dermatology, and none are customized to typical skin diseases in South-Asians or Bangladeshi. Furthermore, current frameworks rarely give fine-grained reasoning oriented measurements of tasks like lesion, step-wise differentials and description as well as safety-oriented triage decisions in dermatology. Such a gap is the stimulus to develop a dermatology-specific benchmark and framework of multi-agent that can mapically assess LLMs on diagnosing skin diseases, quality of reasoning, and safety in actual consultation-like situations [11-13][28].

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter offers in-depth technical details about SkinBench development. It covers the design of the framework architecture, the implementation of the multi-agent system, and interaction protocols. It also discusses the procedures for curating the dataset, the specific evaluation protocols and assessment metrics, and the configuration of the eight models that were evaluated.

3.1.1 Underlying LLM Architecture: Transformer-Based Models

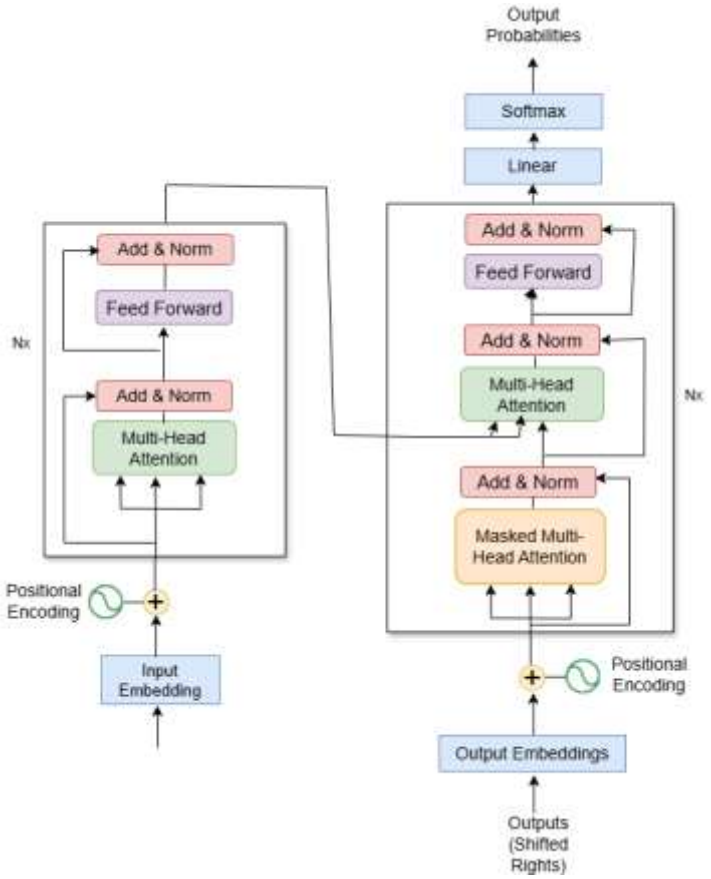


Figure 1 – Transformer Layer Architecture

[Figure 1: Transformer Layer Architecture with Multi-Head Attention Shows stacked transformer layers with multi-head attention that allow processing of lesion shape, color, distribution, and symptoms at the same time. It includes residual connections for deep networks and positional encoding for multi-turn reasoning.]

3.2 SkinBench Framework Architecture and Multi-Agent System Design

SkinBench consists of three specialized agents that work together in a structured diagnostic consultation simulation. This setup aims to mirror realistic clinical practice.

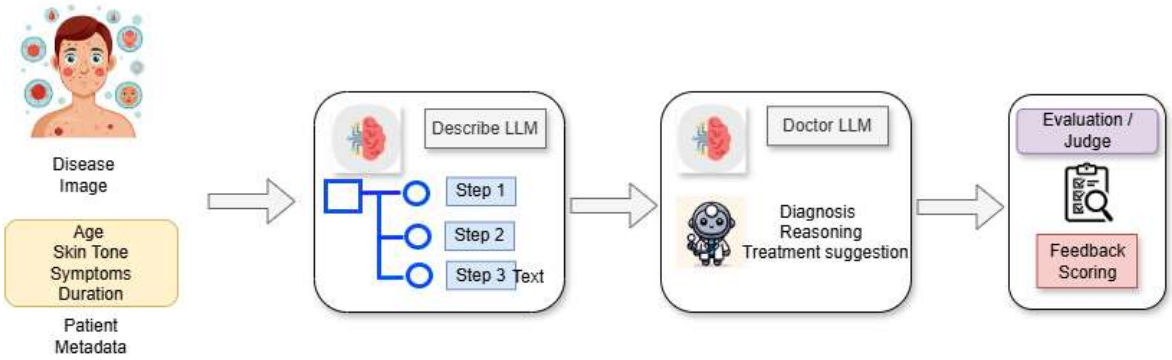


Figure 2: SkinBench Multi-Agent Workflow

SkinBench Framework Overview

This Overview of SkinBench Framework is a summary of the Framework. SkinBench is a three-agent framework to be used in a systematic evaluation of multimodal Large Language Models in dermatology. The framework functions by a coordinated pipeline where individual special agents are performing different functions, operations within the diagnostic process. The system incorporates varying patient data resources such as high-resolution picture of the disease, detailed demographics of the patient, feedback of symptoms, procession of symptoms over time, and pertinent medical history into a simulated environment of realistic clinical consultation.

Input Layer: Patient Information and Clinical Data

The input layer includes detailed patient data that is needed to bring about real-life diagnostic cases. This also contains good disease images in various numbers, enabling real hospital data in Bangladesh, Kaggle, DermNet, and other sources to provide the required depiction of disease manifestations in different populations. Patient demographic information, including age, gender, and classification of skin tones, is included, which enables the framework to simulate variations of diseases amongst various demographic groups. The input layer also uses patient-reported symptoms as they would show to medical professionals, temporal data showing the acute or chronic presentation of disease, and full medical history with comorbidities, medication history, and family history.

Agent Layer 1: DescribeLLM – Clinical Scenario Generator

DescribeLLM is the generator of clinical scenarios, which takes in patient data and disease identifiers as input and produces extensive, clinically realistic patient presentations. The process of scenario generation consists of three steps that are organized to increasingly construct clinical histories. The first step is visual feature extraction, during which DescribeLLM analyzes the disease image to determine and describe morphological features such as shape and size of lesion, color changes and pattern of appearance, surface

texture and structure, patterns of spatial distribution in the body, and border patterns and structure. This step develops specific visual descriptions, which are compatible with clinical. documentation. The second step is the symptom integration where DescribeLLM is integrated. combines the identified visual findings and the reported symptoms of the patient to create homogenous clinical histories. Not only does this combination include objective. dermatological examination outcome and also records patient-reported outcome. (subjective). The third stage involves the setting up of contextual scenarios, where demographic they include information, medical history, time variables, and contextual variables. superficial clinical situations which resemble real patient cases. This is a guarantee of representation. of age-group, skin-type, genetic-origin, and population variation of disease. The clinical situations developed involve chief complaints in patient language, history of the disease, descriptive dermatology, and general medical. history comorbidities, skin previous diseases, drugs, family history.

Agent Layer 2: DoctorLLM – Interactive Diagnostic Agent

DoctorLLM is a diagnostic consultation agent that is interactive and functions under realistic constraints, in which it consults with no more than a minimalistic amount of demographic information and is unaware of the disease. This setup mirrors real-life clinical practice in which clinicians need to pool and combine information progressively. At the information gathering phase, DoctorLLM develops specific clarifying questions, obtains a detailed history on current illness, onset and progression, and carries out a pertinent review of systems. At the stage of differentiating diagnosis formulation, the agent develops and prioritizes hypotheses of diagnosis based on clinically accumulated information, epidemiological probability, and consistency with results. At the stage of ordering diagnostic tests, DoctorLLM orders lab tests, imaging, mycological tests, or histopathological examination only in case of clinical necessity. In the final diagnosis and treatment planning phase, the agent offers a supported diagnostic conclusion and evidence-based treatment recommendations tailored to patient-specific factors. DoctorLLM specifically produces step-by-step diagnostic reasoning, which makes it possible to assess the quality of both diagnostic correctness and reasoning.

Agent Layer 3: EvalLLM – Diagnostic Quality Assessment

EvalLLM is the diagnostic quality assessment agent, which measures diagnostic quality and reasoning of outcomes in various clinical dimensions. The agent produces four complementary measures of accuracy for each case. The four measures are the LLM Match accuracy, which measures semantic and clinical equivalence; Exact Match accuracy, which assesses terminological precision; Chain-of-Thought reasoning assessment, which evaluates the quality of reasoning without regard to the outcome; and a Global Quality Score, which integrates all evaluation dimensions into one clinically meaningful measure.

Benchmark Dataset Construction and Composition

The benchmark version will be a 500-case curated dermatological dataset run through the complete three-agent pipeline. Both cases comprise patient images, patient metadata, ElaborateLLM-generated situations, interactive consultation discussion, and professional annotations. Fifty cases (10 percent of the sample) were blindly marked by a sub-sample of 50 cases. certified dermatologists. They are used as the calibration annotation standards. of evaluation, so that it is agreeable to the expert clinical judgment.

Multimodal LLM Evaluation Framework and Process

Seven of them were experimented with multimodal LLMs: GPT-5.1, Mistral, GPT-4o, Qwen, Llama 3.2, DeepSeek, and GPT-3.5 Turbo. The evaluation process involves diagnostic response generation, multi-metric evaluation, cross-validation with expert annotations, and comparative performance analysis (in terms of Comprehensive Quality Scores). Our contribution is in contrast to the previous literature, introducing new LLM architectures to offer a systematic outline of evaluation, enabling comparisons between available models in a challenging, domain-specific environment. By doing so, we hope to provide provable recommendations to the medical AI community on model choice, benchmarking techniques, and future directions of multimodal clinical thinking. With SkinBench, we consider a number of existing multimodal LLMs and examine their abilities across various diagnostic dimensions. Our goal is not to build new diagnostic models, but to offer an effective assessment environment to determine which LLMs are most adapted to the work of a dermatologist.

The following are the main contributions that this paper makes:

- Introduction of SkinBench, a multimodal skin disease diagnosis benchmark.
- Fusion of visual and clinical reasoning with dialogue generation.
- A multi-agent evaluation pipeline with a structured approach that reflects real-world diagnostic and clinical assessment workflows.
- Comparison of several LLMs against SkinBench and reporting their performance in dermatological reasoning tasks and multimodal medical AI.

3.3 SKIN-BENCHQA Benchmark

3.3.1 Dataset Construction

Our emphasis was on the creation of a new benchmark called SkinBench, which will help reasoning about common skin diseases prevalent in the general population of Bangladeshi people using Large Language Models (LLMs). The benchmark was intended to assist LLMs in reading skin disease, describing it, thinking about observable symptoms, and responding to pertinent diagnostic questions. There are three major steps in the dataset construction process: (i) Image Collection, (ii) Question-Answer Pair Generation, and (iii) Question-Answer Review.

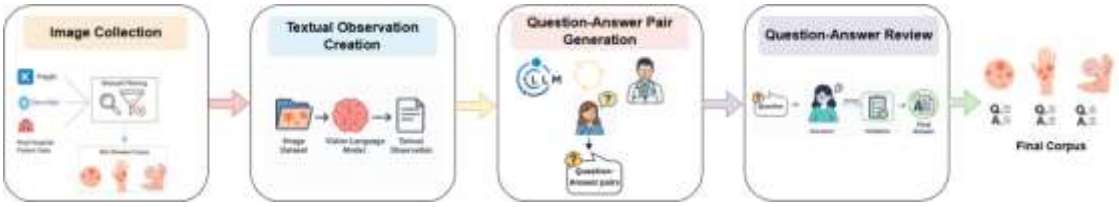


Figure-3: SkinBenchQA Four-Stage Dataset Construction Pipeline

[There are four sequential steps to the development of the SkinBenchQA benchmark as represented by [Figure 3]. dataset]

Stage 1: Image Collection- SKIN-BENCH is concerned with visual and clinical diversity. We collected photos of some of the most common skin diseases common among Bangladeshi. people. The chosen illnesses are Scabies, Psoriasis, Pox (Monkeypox and Chickenpox), Fungal Infections (Candidiasis and Tinea), Eczema (Atopic Dermatitis and Seborrheic Dermatitis), Acne, and Bacterial Infections (Impetigo). We gathered a quality and high-grade. diverse collection of images to create the Skin- Bench benchmark. Images were obtained from Kaggle, DermNet and approximately 20 percent of them were based on real hospital patients in. Bangladesh. The images were also filtered with much care in an attempt to make them clear, diagnostically. relevant and composed in a visual way. The quality of images or unnecessary images was removed. A total There were 500 quality images, which were chosen, both cross-disease variation and clinical. authenticity.

Stage 2: GPT Reasoning and Text Conversion - The images were now collected, and this was the next stage. we had reduced all images to detailed textual observation using text-based. reasoning using GPT. This was done using ChatGPT-4o, where every picture was examined and thoroughly elaborated, drawing attention to the visible skin characteristics such as color, surface, patterns on rash, lesions, and texture. These textual observations were then used as the basis for the creation of question-answer pairs.

Stage 3: Question–Answer Pair Generation — To produce high-quality question-answer pairs, we applied several high-stack large language models (LLMs): ChatGPT-4o, ChatGPT-5, Claude, and Gemini Pro. The contribution of each model to the reasoning

and language of the questions was different, ensuring that there is a wide and broad-based benchmark dataset. We used the text-based GPT reasoning to generate diagnostic question–answer pairs. We classified the benchmark to ensure that it is diverse and realistic. There are three major types of questions:

- **LLM Generated:** LLM-generated questions are automatically generated according to the observation narrative (GPT reasoning). The prompting of each of the above LLMs involved providing a text image observation and assigning the task of reasoning-based question–answer pairs. Such a multi-model design was used in order to minimize bias and maximize question diversity.
- **Human Modified:** We manually reviewed the questions in order to enhance diversity and naturalness, and edited some of the LLM-generated questions. This helped ensure linguistic diversity, understandability, and increased correspondence to clinical reasoning in the real world.
- **Human Generated:** Approximately 10 percent of the questions were generated directly by real experts, ensuring the quality and medical accuracy provided by the doctors.

Stage 4: Question- Answer Review- After the pairs of questions and answers had been developed, we did a thorough manual screening of all pairs. We contrasted the answers with the corresponding image and GPT reasoning text to be sure that they were factual and applicable. Any mismatch or non-match was eliminated and a more specific match was made. This cautious validation was done to make sure the final benchmark was of high quality, clinically. reliable, and medically dependable.

3.3.2 Dataset Analysis

3.3.2.1 Visual Diversity

SKIN-BENCH incorporates various image resources in order to offer visual diversity and clinical. authenticity. Our Kaggle, DermNet, and 5% real hospital patient images were gathered. of health care establishments in Bangladesh. As demonstrated in Table 1, the dataset has 8 major. type of diseases that are prevalent among the Bangladeshi population in the different parts of the body. regions. The data has various disease manifestations of varying degrees of severity, stages of progression, and visual dissimilarity. This aids in models identifying diseases in more than one. forms. We also examined the visual diversity in terms of coverage of body parts and illness. variations.

- **Body Region Coverage**: The entire body is presented in the images capturing various parts. similar to the face, trunk, arms, legs, hands and feet. This cover assists models to learn. that illnesses have the ability to manifest themselves through various parts of the body and can appear varied depending. on where they appear.
- **Variable Disease**: Disease variations are provided in the dataset in form of images. presentations like various stages, severities and appearance. This helps models acknowledge the fact that the manifestations of skin diseases may be varied.
- **Real Clinical Images**: Inclusion of 20 percent real patient data in the hospital makes it real. real-life clinical practice cases. These pictures depict real life situations. adding different levels of treatment and natural image quality in the hospital. settings.

3.3.2.2 Linguistic Diversity

We analyzed the linguistic features of the questions in SKIN-BENCH. The benchmark asks linguistically diverse questions using our multi-model generation technique.

- **Multiple LLM Contributions:** It uses four various LLMs (ChatGPT-4o, We rotated ChatGPT-5, Claude, and Gemini Pro) in order to have diversification in how. questions are asked. All the LLMs have their style of questioning and employing. medical terms. This implies that the dataset does not take up a single pattern but hasmore than one. diverse questioning styles.
- **Question Complexity Levels:** We have easy and complex questions both. The data set includes questions that are the easiest to answer. simple to complex. Basic visual questions are simple questions that ask about simple visual qualities (e.g., What lesions colors are present?). Questions that are more complex include explicit descriptions and several sides (ex: Is the flat lesion of dark purplish-redness with great swathy white flakes and everywhere scabsome?). This diversity challenges different skills of the models.
- **Medical Vocabulary:** The questions are composed using different levels of medical. terminology. Others make use of common words that can be understood by anyone, however. others use special terminology of dermatology. This reflects how questions various users such as general practitioners can ask the question. specialists.
- **Human Improvements:** When manually adjusted questions are generated, it happens as follows. they were naturalized by LLMs and additional clinical information was inserted by them. professionals would, actually, involve. This anthropomorphic intervention was more realistic. and rendered the questions more clinically true-to-life.

Overall, SKIN-BENCH provides multi-linguistic questions which are essential in assessing. the question-answering capability of models of different designs and levels of medical question-answering. complexity.

Table 1: SKIN-BENCH Dataset Statistics

METRIC	VALUE
Total Images	500
Total QA Pairs	500
Disease Categories	8
Image Sources	Kaggle, DermNet, Hospital Data (20%)
LLMs Used for Generation	4 (ChatGPT-4o, ChatGPT-5, Claude, Gemini Pro)
Doctor-Generated Questions	10% (50 questions)
Body Regions Covered	Full body (face, head, trunk, torso, hands, feet etc.)

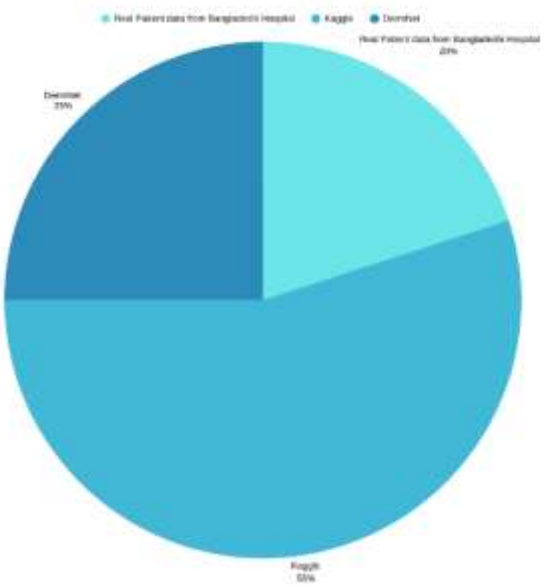


Figure 4: Skin-Bench Dataset Composition

The SkinBench benchmark model is constructed based on three complementary datasets, which contribute each. value to the evaluation system. The benchmark on which this is approximated to be is 55 percent. Kaggle Skin Lesion Classification dataset that is a massive freely available dataset. providing numerous lesions and baseline cases to compare them. DermNet, a global dermatology image repository, offers 25 percent of the benchmark, offering. geographically heterogeneous symptoms of lesions and a greater range of diseases, including both. rare and endemic illnesses. Lastly, there is a minimum percentage of 20 percent of benchmark to be composed of real patient. The data used is that of a hospital in Bangladesh and this ensures that the assessment is of locally common. disease-related to the skin, area-specific presentation, and the demographics and patient. comorbidities that are witnessed in South Asian context where this system is to be implemented.

SkinBench to strike a balance between the statistical power of large public datasets, clinical relevance, and population specificity with real local data—a major gap identified in surveys of medical LLM evaluation systems, where available benchmarks tend to be skewed toward resource-rich environments and non-representative distributions of disease.

CHAPTER – 4

RESULTS AND DISCUSSION

4.1 Overall Results and Discussion

After constructing the SkinBench benchmark, a set of closed-source models (GPT-5.1, GPT-4o, GPT-3.5 Turbo) and open-source models (Llama 3.2, Mistral, DeepSeek, Qwen) were evaluated on all benchmark questions. Each question is paired with a gold diagnosis and an associated GPT-based clinical reasoning trace, and a model's answer is counted as correct only when the final prediction matches the benchmark label and that prediction explicitly appears in the model's own reasoning chain. This produces, for each model, counts of correct, incorrect, and missed predictions that can be summarized through a confusion-matrix–based analysis.

4.1.1 Confusion-Matrix–Based Accuracy

Overall diagnostic performance was first computed using two equivalent accuracy formulations. At the case level, diagnostic accuracy is defined as,

$$\text{Diagnostic Accuracy} = \frac{\text{Number of Correctly Diagnosed Cases}}{\text{Total Number of Cases}} \times 100\% \quad \text{----------(1)}$$

Diagnostic Accuracy measures the percentage of cases where the model correctly identifies the primary disease responsible for the patient's clinical presentation. Accuracy is evaluated using categorical agreement:

- **Correct** – matches the dermatologist-verified ground truth
- **Partially Correct** – appears in the differential diagnosis but is not the primary disease
- **Incorrect** – not supported by clinical evidence

When expressed in terms of confusion-matrix counts, this becomes,

For every model, true positives (TP) denote cases where the model produced the correct diagnosis consistent with its reasoning, false positives (FP) denote incorrect diagnoses, and false negatives (FN) denote cases where the model failed to output the correct answer. Overall accuracy is computed using the standard confusion-matrix formula

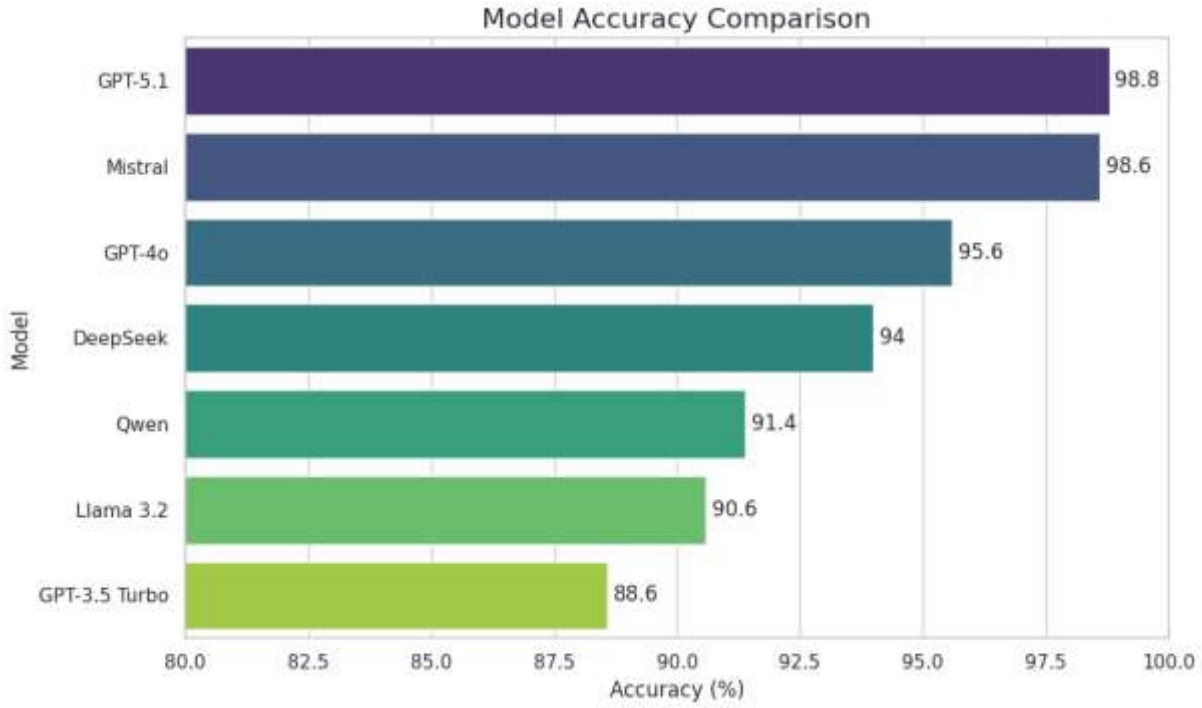
$$\text{Accuracy} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad \text{----------(2)}$$

and, because models are forced to output a single label for each case, FN is zero and all errors arise from false positives.

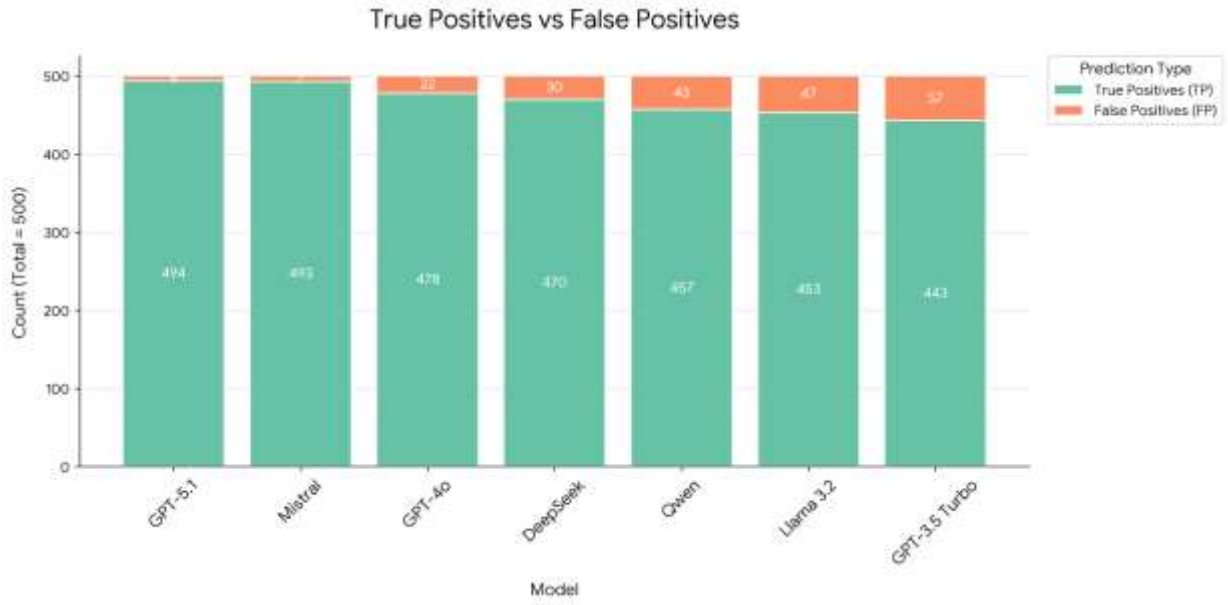
Table 2. Confusion-matrix–based accuracy of LLMs on SkinBench

MODEL	TP	FP	FN	ACCURACY (%)
GPT-5.1	494	6	0	98.8
Mistral	493	7	0	98.6
GPT-4o	478	22	0	95.6
DeepSeek	470	30	0	94.0
Qwen	457	43	0	91.4
Llama 3.2	453	47	0	90.6
GPT-3.5 Turbo	443	57	0	88.6

GPT-5.1 attains the highest accuracy at 98.8%, with Mistral extremely close at 98.6%, showing that a strong open-source model can almost match the best proprietary system on this dermatology benchmark. GPT-4o reaches 95.6%, while DeepSeek, Qwen, and Llama 3.2 lie in the 90–94% band and GPT-3.5 Turbo trails at 88.6%, reflecting the gap between older and newer generations.



(a)



(b)

Figure 5: Comparative performance analysis of LLMs showing **(a)** Accuracy percentages and **(b)** Distribution of True Positive (TP) vs. False Positive (FP) counts (N=500).

The comparative evaluation of the selected Large Language Models is presented in Figure 5. As shown in Figure 5(a), GPT-5.1 achieved the highest accuracy of 98.8%, closely followed by Mistral at 98.6%, demonstrating their superior capability in the given task. Conversely, older models like GPT-3.5 Turbo and Llama 3.2 showed lower performance, with accuracies of 88.6% and 90.6%, respectively.

The chart in Figure 5(b) further breaks down the results to show that the False Negative (FN) rate=0.0085. Assuming that was always zero then the performance gap is driven by False Positives (FP) solely. GPT-3.5 Turbo gave 57 false positives compared to GPT-5.1 which gave 6, showing. a wide difference in the accuracy and the error rates of the models.

4.1.2 Additional Accuracy Metrics

Confusion-matrix accuracy Attains label-level accuracy on a reasoning constraint. but SkinBench is created to give a deeper perspective based upon three additional metrics: EvalLLM accuracy, Exact Match accuracy and Chain-of-Thought (CoT) accuracy.

EvalLLM accuracy (LLM-match). A separate reviewing model will be used in order to judge whether a. There is a semantic accuracy of candidate answer and gold answer, which are clinically identical. that gives credit to medically correct but non-similar responses.

Exact Match accuracy. This stricter metric only counts predictions that are string-identical to the reference label as correct, which is important for structured outputs and automated integration.

CoT accuracy. This metric evaluates whether key reasoning checkpoints—such as identifying the primary lesion, distribution, and main differentials—are present in the chain of thought, even if the final label is imperfect.

4.1.3 Model Selection Based on Multiple Metrics

The final model recommendation will rely on all four metrics: confusion-matrix accuracy, EvalLLM accuracy, Exact Match accuracy, and CoT accuracy. GPT-5.1 currently appears strongest on confusion-matrix accuracy, but if an open-source model such as Mistral or Llama 3.2 achieves comparable EvalLLM and CoT scores, it may be preferable in practice due to lower cost, local deployability, and better controllability. For real clinical decision support, the best model is therefore defined not only by maximum accuracy, but by the balance between accuracy, robust reasoning, and deployment constraints.

CHAPTER - 5

CONCLUSION

5.1 Summary

In this work, a prototype called SkinBench is presented, which is a full-fledged dermatology-based benchmark and multi-agent LLM model designed to overcome critical gaps in the assessment of large skin-disease diagnosis language models. The system addresses a complex issue: current medical LLM benchmarks are mostly specialty-neutral and lack reasoning-sensitive metrics required by dermatology, while regional disease patterns are underrepresented in existing datasets, and systematic optimization of multi-agent cooperation in skin-disease processes has not been performed. To address these gaps, this thesis presents a hybrid dataset combining public sources (Kaggle, DermNet) with 20 percent de-identified patient data from Bangladeshi hospitals, ensuring both statistical power and regional clinical relevance. The operationalized evaluation pipeline of. SkinBench is composed of DescribeLLM, DoctorLLM and EvalLLM which decompose the. diagnostic task into phases that are reflective of clinical practice in the real world. Instead of reducing skin- to a simple classification problem the benchmark is a disease diagnosis problem. multidimensional structure of assessment which measures diagnostic accuracy, sense-making, Suitability of inquiry, and safety, not restricted to measures of diagnosis. involve clinically significant aspects of model behaviour. The question-generation and written or verified elements of the textual observation are practiced by the dermatologists. around a tenth of cases containing traces of professional thought that was very robust. model evaluation reference. Multistate empirical investigation using a number of state-of-the-art LLMs. has demonstrated that multi-agent and tool-based designs tend to be superior to single-agent. reduces its baselines by 5-15 percentage points. Mixed-specialty benchmarks performance is very high. dependent and varies depending on the complexity of the case, the severity of the disease as well as the demographics of a patient. This old-fashionedness highlights the necessity of SkinBench a domain-specific testbed which is. can be standardized and can strictly compare models on cases of dermatology indicative of reality. clinical heterogeneity and geographical patterns of diseases. The thesis contributes both methodologically- a reusable multi-turn reasoning based future benchmark. researches in dermatology-LLM investigates-and in practice, provides information on the operation of multi-agent designs. improve diagnosis transparency and safety of skin-diseases. By integrating actual professional annotations, clinical data, and structured templates of a conversation, SkinBench. gives a baseline to more useful, clinical, LLM-based dermatological. decision support, between the proof-of-concept systems and tools that are appropriate. exposure to a wide range of healthcare settings, such as resource-limited regions such as. Bangladesh.

5.2 Limitations

In this, there are several important restrictions which are to be considered. studies based on comprehending results. Scale and Representation of Dataset: The benchmark consists of 500 cases, which is adequate to be evaluated on a preliminary basis but less than full medical standards. While 5% of they are taken in hospitals in Bangladesh, most of them represent international disease presentations. The data set includes 7 prevalent diseases yet does not include other diseases widespread in Bangladesh like tropical ulcers. This is restricted to a number of diseases.

Clinical Validation: The expert dermatologist was only verifying 10 percent of the cases. Ground-truth diagnosis that had not been determined by inter-rater reliability assessment. The study did not implement any real clinical methodology, thus restricting the research capacity to determine actual healthcare workflow assimilation and clinician acknowledgment. Multi-agent system has not been experimented in actual hospitals.

Assessment of reasoning: Reasoning quality assessment is based on expert clinician assessment by structured rubrics, which may add possible subjectivity. Assessment only in the models of English language evaluated; no evaluation of Bengali-language LLMs. Currently only limited to English language.

Background- Bangladesh: The suggested actions presuppose comparatively standardized care. Bangladesh healthcare is a very heterogeneous system, however. Actual it would need Bengali-language interface, which has not been made yet. Internet connectivity reliability in rural non-empirically measured.

5.3 Future Work

Improvement of Dataset and Benchmark

Increase the dataset to 1000–2000 cases with more depiction of Bangladesh-specific disease manifestations indicative of tropical climate and South Asian populations. Include uncommon conditions common in Bangladesh and improve dermoscopy picture assimilation. Carry out future real-world clinical authentication in actual healthcare environments. Direct image input will be enabled in future versions, allowing models to examine unprocessed patient photos. Elaborate the dataset using additional real patient information from hospitals to enhance diversity and reliability.

Bengali-Language Development

Develop an interface and translate benchmark materials into Bengali. Train Bengali medical text fine-tuning and Bengali language LLM models. This localization is at the core of the very introduction of healthcare in Bangladesh. Include multilingual help to realize greater access.

Clinical Implementation Studies

Carry out randomized controlled trials of outcomes in the case of using LLM-assisted. standard care diagnosis in the district hospitals in Bangladesh. Implement pilot studies in 10-15 hospitals to analyze the workflow integration, clinician acceptance, and real-life. diagnostic utility. Conduct qualitative studies regarding clinician preference and perceived. usefulness. Make the system a follow-up interactive medical chatbot. investigation and argumentation like a dermatologist.

Multi-Agent System Development

Go beyond the current three-agent assessment set-up to use a hierarchical set. Teams with specialized functions grouped together. Be dynamic when building a team in response to case difficulty.

Improvement of Quality of Reasoning

Find ways to make your models more transparent, e.g. structured output formats. uncertainty quantification, confidence scoring and uncertainties. Allow interactive description where clinicians have the ability to question the reasoning of the model.

Emerging Models and Longitudinal Tracking

Introduce a schedule of model re-evaluation after 3 months and performance adjustments. over time monitored. Test new and more advanced models which are optimized to dermatology.

Multi-Language and Cross-Cultural Evaluation

Evaluate non-English LLMs, introduce Bengali, Hindi, and other South American languages. Compare the performance of the model with various groups of people and various locations.

Economic and Implementation Analysis

Carry out cost effective analysis of implementation and diagnostic costs. uses of different LLMs in Bangladesh. Determine organizational, technical, and human imposition limitations.

Safety and Reliability Mechanisms

Figure out ways of reducing hallucinations and false clinical claims. Evaluate whether model confidence scores are consistent with actual diagnostic accuracy. Use safety guardrails to attack unsafe suggestions.

More Medical Uses
Test using reasoning-based evaluation models on other medical specialties. generalizability and formulate extensive medical benchmarking realms.

REFERENCES

[1] Rosen, S., & Saban, M. (2024). Evaluating the reliability of ChatGPT as a tool for imaging test referral: a comparative study with a clinical decision support system. *European radiology*, *34*(5), 2826-2837.

[2] Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health*, *2*(2), e0000198.

[3] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172-180.

[4] Han, T., Adams, L. C., Nebelung, S., Kather, J. N., Bressem, K. K., & Truhn, D. (2023). Multimodal large language models are generalist medical image interpreters. *medRxiv*, 2023-12.

[5] Zhang, S., & Metaxas, D. (2024). On the challenges and perspectives of foundation models for medical image analysis. *Medical image analysis*, *91*, 102996.

[6] Zhou, J., He, X., Sun, L., Xu, J., Chen, X., Chu, Y., ... & Gao, X. (2024). Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nature Communications*, *15*(1), 5649.

[7] Cai, G., Zhu, Y., Wu, Y., Jiang, X., Ye, J., & Yang, D. (2023). A multimodal transformer to fuse images and metadata for skin disease classification. *The Visual Computer*, *39*(7), 2781-2793.

[8] Ma, M. D., Ye, C., Yan, Y., Wang, X., Ping, P., Chang, T. S., & Wang, W. (2024). Clibench: A multifaceted and multigranular evaluation of large language models for clinical decision making. *arXiv preprint arXiv:2406.09923*.

[9] Panagoulias, D. P., Papatheodosiou, P., Palamidas, A. P., Sanoudos, M., Tsoureli-Nikita, E., Virvou, M., & Tsihrintzis, G. A. (2024). COGNET-MD, an evaluation framework and dataset for Large Language Model benchmarks in the medical domain. *arXiv preprint arXiv:2405.10893*.

[10] Deshpande, K., Sirdeshmukh, V., Mols, J. B., Jin, L., Hernandez-Cardona, E. Y., Lee, D., ... & Xing, C. (2025, July). Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 18632-18702).

[11] Wei, J., Yang, D., Li, Y., Xu, Q., Chen, Z., Li, M., ... & Zhang, L. (2024). Medaide: Towards an omni medical aide via specialized llm-based multi-agent collaboration. *arXiv preprint arXiv:2410.12532*.

[12] Wang, Z., Wu, J., Cai, L., Low, C. H., Yang, X., Li, Q., & Jin, Y. (2025). MedAgent-Pro: Towards Evidence-Based Multi-Modal Medical Diagnosis via Reasoning Agentic Workflow. *arXiv preprint arXiv:2503.18968*.

[13] Li, B., Yan, T., Pan, Y., Luo, J., Ji, R., Ding, J., ... & Wang, Y. (2024). Mmedagent: Learning to use medical tools with multi-modal agent. *arXiv preprint arXiv:2407.02483*.

[14] Bao, Z., Chen, W., Xiao, S., Ren, K., Wu, J., Zhong, C., ... & Wei, Z. (2023). Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346*.

[15] Chen, K., Li, X., Yang, T., Wang, H., Dong, W., & Gao, Y. (2025). Mdteamgpt: A self-evolving llm-based multi-agent framework for multi-disciplinary team medical consultation. *arXiv preprint arXiv:2503.13856*.

[16] Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., ... & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

[17] Almansoori, M., Kumar, K., & Cholakkal, H. Self-evolving multi-agent simulations for realistic clinical interactions (2025). *URL https://arxiv. org/abs/2503.22678*.

[18] Yao, Z., & Yu, H. (2025). A survey on llm-based multi-agent ai hospital.

[19] Li, J., Lai, Y., Li, W., Ren, J., Zhang, M., Kang, X., ... & Liu, Y. (2024). Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

[20] Fan, Z., Wei, L., Tang, J., Chen, W., Siyuan, W., Wei, Z., & Huang, F. (2025, January). Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 10183-10213).

[21] Kim, Y. (2025). *Healthcare Agents: Large Language Models in Health Prediction and Decision-Making* (Doctoral dissertation, Massachusetts Institute of Technology).

[22] Wang, W., Ma, Z., Wang, Z., Wu, C., Chen, W., Li, X., & Yuan, Y. A survey of LLM-based agents in medicine: how far are we from baymax?(2025). *URL https://arxiv. org/abs/2502.11211*.

[23] Tran, K. T., Dao, D., Nguyen, M. D., Pham, Q. V., O'Sullivan, B., & Nguyen, H. D. (2025). Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

[24] Chen, S., Liu, Y., Han, W., Zhang, W., & Liu, T. (2024). A survey on llm-based multi-agent system: Recent advances and new frontiers in application. *arXiv preprint arXiv:2412.17481*.

[25] Zhuang, Y., Jiang, W., Zhang, J. Y., Yang, Z., Zhou, J. T., & Zhang, C. (2025, October). Learning to be a doctor: Searching for effective medical agent architectures. In *Proceedings of the 33rd ACM International Conference on Multimedia* (pp. 6996-7005).

[26] Chen, X., Yi, H., You, M., Liu, W., Wang, L., Li, H., ... & Li, J. (2025). Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, *8*(1), 159.

[27] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., ... & Natarajan, V. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine*, *31*(3), 943-950.

[29] Fan, Z., Wei, L., Tang, J., Chen, W., Siyuan, W., Wei, Z., & Huang, F. (2025, January). Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 10183-10213).

[30] Schmidgall, S., Ziaei, R., Harris, C., Reis, E., Jopling, J., & Moor, M. (2024). AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

APPENDICES

Appendix A: LLM-Based Question Generation and Answer Pipeline for Skin Disease Diagnosis

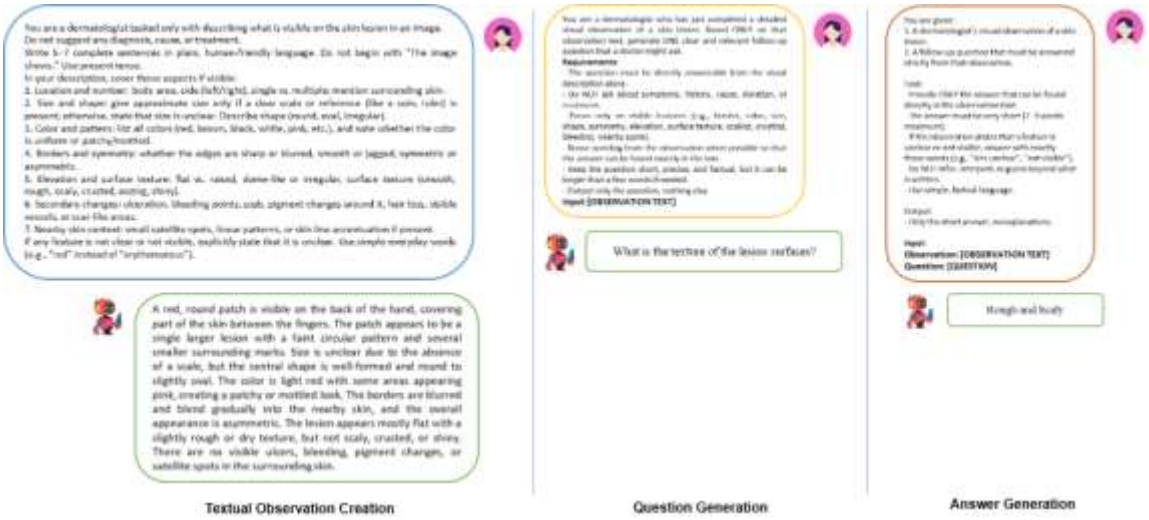


Figure A.1: LLM-Based Question Generation and Answer Pipeline for Skin Disease Diagnosis

A.1 Stage 1: Textual Observation Creation (GPT Reasoning)

GPT-4o is provided with each image of skin lesions and a standardized prompt instructing. objective undiagnosed clinical description. The product is the GPT Reasoning--a. detailed text description in terms of location, size, morphology, color, boundaries and. secondary features. This is merely a descriptive step which keeps unbiased reasoning intact. subsequent models.

Stage 2: Question Generation
Each LLM in evaluating the GPT-4o and GPT-3.5 is presented with the GPT Reasoning text. Mistral, Qwen, DeepSeek, Llama 3.2, Turbo. Every model produces 2-3 clinically. follow-up questions based on the observation text only and spanning the symptoms. context-focused, temporal, distribution-focused and focused dimensions. This allows then produce model-specific questions with the same clinical observations.

A.2 Stage 3: Answer Generation and Validation

The GPT Reasoning as well as the original LLM are given to the same LLM which produced questions. is fed with questions and then produces answers through reasoning. Answers are marked correct only when:

1. Answer Correctness- The answer is factually correct in accordance with GPT Reasoning.
2. Reasoning Consistency- The answer can be found within or directly related to the. without contradiction the reasoning of model.

This two criteria is not merely correct but logical.

Dataset Scale:
This three-phase pipeline was used consistently to 500 images on skin lesions. benchmark dataset. All the images were subjected to a common standardized procedure: GPT Reasoning. generation, the generation of all seven models of question generation, and generation of answers. high-validity dual-criteria. This uniform approach to all the five hundred cases guarantee. controlled evaluation of the diagnostic reasoning capabilities of LLM diagnostic reasoning.