

Cultivating Multimodal Intelligence: Interpretive Reasoning and Agentic RAG Approaches to Dermatological Diagnosis

Notebook for the ImageCLEF MEDIQA-MAGIC Lab at CLEF 2025

Karishma Thakrar^{1,*†}, Shreyas Basavatia^{1,†} and Akshay Daftardar^{1,†}

¹Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332

Abstract

The second edition of the 2025 ImageCLEF MEDIQA-MAGIC challenge [1], co-organized by researchers from Microsoft, Stanford University, and the Hospital Clinic of Barcelona, focuses on multimodal dermatology question answering and segmentation, using real-world patient queries and images. This work addresses the Closed Visual Question Answering (CVQA) task, where the goal is to select the correct answer to multiple-choice clinical questions based on both user-submitted images and accompanying symptom descriptions.

This task presents several challenges: consumer health questions are often noisy, imprecise, and often lack relevant clinical context. Unlike real-world settings where physicians can iteratively ask follow-up questions, the system must make medical decisions with high accuracy based on a single multimodal static patient interaction. This raises the difficulty of building models that generalize well and increases the risk of clinically significant misclassifications.

To address these limitations, not only predictive accuracy, but also reasoning ability and explainability were prioritized. The proposed approach combines three core components: (1) fine-tuning open-source multimodal models from the Qwen, Gemma, and LLaMA families on the competition dataset, (2) introducing a structured reasoning layer that reconciles and adjudicates between candidate model outputs, and (3) incorporating agentic retrieval-augmented generation (agentic RAG), which adds relevant information from the American Academy of Dermatology's symptom and condition database to fill in gaps in patient context.

The team achieved second place with a submission that scored sixth, demonstrating competitive performance and high accuracy. Beyond competitive benchmarks, this research addresses a practical challenge in telemedicine: diagnostic decisions must often be made asynchronously, with limited input and with high accuracy and interpretability. By emulating the systematic reasoning patterns employed by dermatologists when evaluating skin conditions, this architecture provided a pathway toward more reliable automated diagnostic support systems.

Keywords

Agentic Vision-Language Models, Generative AI for Clinical Decision Support, Medical Natural Language Understanding, Multimodal Healthcare Intelligence and Reasoning Systems, Knowledge-Grounded Medical Question Answering, Ensemble Learning for Telemedical Diagnostics, Medical Domain Specific Model Adaptation

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

✉ karishma.thakrar@gatech.edu (K. Thakrar); sbasavatia3@gatech.edu (S. Basavatia); adaftardar3@gatech.edu (A. Daftardar)

🌐 <https://github.com/karishmathakrar> (K. Thakrar); <https://github.com/Basavatia-Shreyas> (S. Basavatia); <https://github.com/akshaydaf> (A. Daftardar)

🆔 0009-0008-2563-7370 (K. Thakrar); 0009-0000-6467-802X (S. Basavatia); 0009-0006-5578-5783 (A. Daftardar)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²Project code available at: <https://github.com/karishmathakrar/arc-mediqa-magic-2025>

1. Introduction

Dermatological diagnostics present complex challenges that require integrating diverse data types including visual images, patient narratives, and contextual information. Traditional diagnostic approaches rely on in-person examinations to interpret subtle visual features and discuss reported symptoms along with patient history, a process that doesn't seamlessly translate to remote consultation settings where information and interactions can be both limited and unclear. More specifically, telemedicine consultations introduce several constraints: variable image quality from consumer devices, imprecise symptom descriptions from patients unfamiliar with medical terminology, and limited opportunity for follow-up questions. These factors make accurate diagnosis in remote settings substantially more challenging.

The second edition of the ImageCLEF MEDIQA-MAGIC challenge [1], co-organized by researchers from Microsoft, Stanford University, and Hospital Clinic of Barcelona, addresses telemedicine constraints through the DermaVQA dataset [2] that seeks to emulate real-world dermatological consultations. It combines patient-captured images with accompanying clinical text and presents two key tasks: (1) segmenting regions of interest in dermatological images and (2) answering closed-ended multiple-choice questions ("Closed Visual Question Answering" or "CVQA") about the conditions shown. These questions systematically categorize dermatological presentations across dimensions including affected body areas, lesion characteristics (size, texture, color, count), symptom duration, and associated sensations like itching. By structuring the challenge around real consumer health queries and standardized question schemas developed by certified dermatologists, the competition creates a testbed for multimodal systems tasked with making accurate diagnostic assessments despite limited training examples, incomplete coverage, and sparse contextual information, mirroring the practical constraints of asynchronous telemedicine.

In this paper, we propose methods to advance dermatological diagnosis, including model fine-tuning, integrative clinical reasoning, and agentic retrieval-augmented generation. Our methodology, while aimed at improving performance on ambiguous cases, contributes to the broader goal to enable transparency, giving clinicians insight into how the system arrived at its conclusions. Unlike traditional extraction-based techniques such as pattern mining and topic modeling [3], our system is designed to emulate the cognitive processes and diagnostic reasoning of experienced dermatologists when faced with incomplete or ambiguous data. Our key contributions are:

1. **Fine-tuning multimodal models on dermatology data to simulate domain-specific clinical training;**
2. **Integrating predictions across models to emulate clinicians' deliberation during differential diagnosis;**
3. **Introducing an agentic retrieval system that reflects, revises, and queries external knowledge—mirroring how clinicians consult references and reconsider diagnoses under uncertainty.**

These components were implemented using seven open-source vision-language models (VLMs) alongside Google's multimodal Gemini 2.5 Flash, which served as the reasoning engine in our agentic RAG and explanation layers. Our system achieved second place in the MEDIQA-MAGIC 2025 challenge with an accuracy of 0.71, outperforming the overall team average of 0.59. This highlights the effectiveness of our integrated strategy.

Large language models often rely on black-box reasoning, which poses a major challenge for clinical adoption. In high-stakes settings like dermatology, even accurate models may be met with skepticism if their decision-making process is opaque. Our system helps bridge this gap by providing explainable, context-aware responses tailored to clinicians. This improves trust in remote dermatological diagnosis and contributes more broadly to the development of safer, more reliable AI tools for telehealth and clinical decision support.

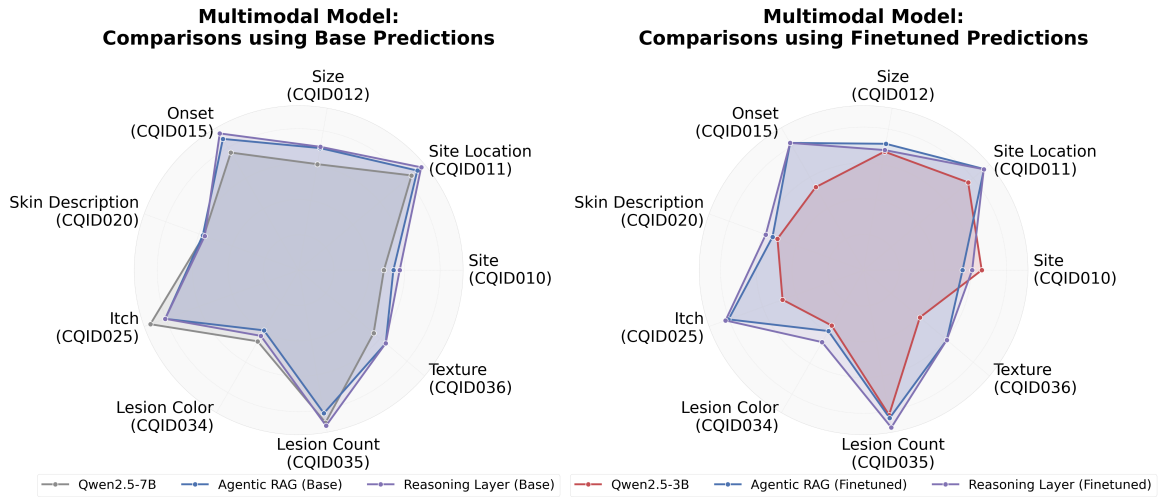


Figure 1: Comparative Analysis of Top-performing Models. Best-performing baseline and fine-tuned models are compared with Agentic RAG and Reasoning Layer enhancements, demonstrating comparable accuracy gains in both cases from Agentic RAG and the Reasoning Layer.

2. Related Work

Multimodal diagnostic systems have advanced medical AI by combining vision and language understanding, with growing applications in dermatology. SkinGPT-4 [4] pioneered interactive diagnosis by aligning vision transformers with LLaMA-2-13b-chat for skin evaluation, while Med-Gemini [5] achieved substantial performance gains through fine-tuning on medical data. Specialized reasoning systems like MedCoT [6] introduced hierarchical expert verification frameworks and approaches like Cross-Attentive Fusion [7] leveraged segmentation models for diagnostic reasoning in controlled clinical settings. While these systems demonstrated the potential of large-scale models through broad medical fine-tuning, they mostly optimized individual model components rather than reasoning across multiple predictions to simulate deliberative clinical judgment.

Building on these advancements, foundation model innovations have further expanded multimodal understanding through diverse approaches. M²Chat [8] balances visual and semantic features with learnable gating mechanisms, and LLM2CLIP [9] enhances vision-language alignment through contrastive fine-tuning. Quality control has evolved with Label Critic [10] automatically assessing medical annotations through anatomical knowledge. However, these architectural and data improvements primarily targeted performance optimization rather than addressing the unique challenges of reasoning under uncertainty with incomplete information.

Retrieval-augmented generation has shown promise for addressing standalone language model limi-

tations in healthcare. Evaluations [11] demonstrated RAG systems can improve medical QA accuracy by up to 18% over chain-of-thought prompting in structured question-answering tasks, with benchmarking revealing optimal retriever-knowledge source combinations. Advanced healthcare RAG frameworks [12] incorporated rationale-guided retrieval and balanced corpus sampling to mitigate bias. Yet existing RAG systems often treated retrieval and generation as isolated steps, limiting their ability to reason across multiple retrieved perspectives and selectively augment missing context for faithful clinical inference under real-world constraints.

Despite progress in medical visual question answering for structured clinical data, significant gaps persist for telemedicine applications where patient-submitted images are informal, incomplete, and suboptimally captured. Several approaches targeted controlled clinical-grade imaging rather than the noisy, heterogeneous data typical of remote consultations [13, 14]. While recent work [15] advanced interpretability through concept extraction on standard benchmarks, approaches systematically addressing representation gaps between clinical and consumer imaging through external knowledge integration remain limited. This work addressed these limitations by developing a reasoning-focused system combining contextual retrieval with ensemble-like decision-making, specifically designed for the information asymmetries inherent in patient-provided medical context for remote dermatological consultations.

3. Exploratory Data Analysis

The DermaVQA dataset [2] used in the competition was organized around 300 unique patient encounters, each representing a complete dermatological case. Every encounter included patient-level context (query titles and clinical descriptions), multiple dermatological images (about three per case), and structured diagnostic annotations spanning 27 questions grouped into 9 major clinical question families. A variety of key assessment domains were covered such as body coverage extent (CQID010), anatomical location (CQID011), lesion size (CQID012), temporal onset (CQID015), morphology (CQID020), symptomatology (CQID025), and lesion attributes like color, quantity, and texture (CQID034–036). Several of these included multiple slots to capture cases involving lesions with diverse locations or characteristics, resulting in 2,700 total encounter-question family combinations. Of these, only 6.56% (177/2,700) had multiple valid answers, primarily in CQID011 (31% of encounters), CQID020 (25%), and CQID012 (3%).

Analyzing the dataset revealed significant class imbalances and consistent trends across clinical question types. The frequency of "Not mentioned" responses varied by domain, entirely absent in core assessments like body coverage (CQID010) but common in more subjective areas such as lesion texture (CQID036: 56.0%) and itch (CQID025: 55.0%). Among non-default answers, skewed distributions were observed, reflecting both clinical prevalence and potential annotation bias. For anatomical location (CQID011), extremities were most frequently reported (upper: 31.3%, lower: 23.6%), while head (10.1%), neck (5.2%), back (9.7%), and other locations (6.3%) were less common. In lesion morphology (CQID020), "raised or bumpy" was the most common label (39.2%), followed by thin or close to the surface (12.7%) and crust lesions (11.6%). Lesion quantity (CQID035) responses were heavily skewed toward multiple lesions (81.3%) over single presentations (16.0%). There was also substantial variation across annotators. One annotator used "Not mentioned" 77.8% of the time, while others mostly ranged from 0% to 33%. These inconsistencies in annotation style were difficult to reconcile within a single model, likely introducing error into the final results

3.1. Dataset Limitations

While DermaVQA represents a valuable and novel contribution to multimodal clinical NLP, several real-world characteristics of the dataset posed challenges for both modeling and evaluation. Many questions featured semantically overlapping or clinically ambiguous answer choices that reflect genuine diagnostic uncertainty. For instance, in CQID034, options like “red” and “pink” or “white” and “hypopigmentation” often appear visually similar in dermatological images, yet annotators were required to select only one, forcing distinctions that may not align with ground truth. CQID020 exhibited similar overlap: labels such as “raised or bumpy,” “thick or raised,” and “wart” were rarely co-selected, only 3 out of 180 encounters included more than one, despite their non-mutually-exclusive nature.

Likewise, clinically co-occurring features like “scab” and “weeping” appeared together in just 3 of 40 relevant cases. This suggests annotators often chose a single, representative label to avoid redundancy, even for questions with multiple valid answers possible. Additionally, overlapping content across question IDs introduced further ambiguity: CQID010 asked about the extent of affected areas with options like “limited area” and “widespread,” while CQID012 asked about lesion size using nearly identical descriptors such as “larger area” and “size of palm.” These patterns reflect real diagnostic ambiguity and inter-observer variability, likely contributing to annotation bias and label inconsistency that complicate both training and evaluation.

Beyond content-level ambiguity, structural issues in the dataset made modeling more difficult. Patient contexts were user-generated and not standardized, often lacking essential clinical details, containing informal or ungrammatical language, or embedding implicit questions. For example, the context in ENC00002 included: “What is on the bottom of the right foot?” Some question formats were also misaligned with classification-style modeling: CQID011 included “other (please specify),” CQID034 had “combination (please specify),” and CQID012 embedded ambiguity directly into the prompt itself with “How large are the affected areas? Please specify which affected area for each selection.” In total, these formulations encouraged open-ended answers that conflicted with the discrete-label format used in evaluation.

Despite efforts in preprocessing and prompt design to mitigate these challenges, deeper inconsistencies remained. Some ground truth labels directly contradicted the available context or visual evidence. In ENC00023, for example, “upper extremities” was annotated as an affected area, though this was not mentioned in the clinical data. Similarly, the physician reported “no itching,” despite visible scratch marks in the images. In other cases, symptoms like itch may not have been reported at all, as the data, sourced from Reddit, was not structured for clinical completeness. An additional example of the data format along with an analysis of the ground truth labels has been provided in Table 3 (see Appendix). These contradictions highlight a central challenge in multimodal reasoning: reconciling conflicting or incomplete inputs across modalities. Addressing such gaps required higher-level reasoning strategies that could go beyond surface cues. These issues were further compounded by variability in image quality, from sharp clinical photos to blurry, user-submitted images, and demographic or platform-related disparities across subsets (i.e., IIFY vs. Reddit), which affected language complexity, answer distributions, and skin tone representation.

To address these challenges, we adopt a fully generative approach, using LLMs with augmented reasoning or agentic RAG to select from predefined answer choices rather than rigidly mapping inputs to fixed labels. This flexibility enables the model to navigate ambiguity, incomplete information, and variability common in real-world patient descriptions. While this may reduce alignment with underspecified or noisy ground truth labels under traditional metrics, the resulting outputs are more

clinically meaningful, interpretable, and better aligned with real-world diagnostic reasoning.

4. Data Preprocessing

The preprocessing pipeline integrated data from three primary sources across the train, validation, and test splits:

- `[split].json` — containing metadata for each clinical encounter, including associated image identifiers;
- `[split]_cvqa.json` — containing answer annotations (as indices) for 27 diagnostic questions per encounter;
- `closedquestions_definitions_imageclef2025.json` — providing question text, answer options, and metadata shared across splits.

Although questions were annotated at the encounter level, each associated image was treated as an independent sample during preprocessing. The pipeline separated out individual images per encounter and paired them with the same unified prompt, creating one row per image-question pair. This allowed each image to be processed independently during individual model inference, while still linking back to the same encounter-level supervision.

Answer indices were mapped to their corresponding textual labels using the provided answer options. For multi-slot question families (i.e., CQID034-A to CQID034-F), we grouped responses by their shared base ID (i.e., CQID034), which introduced a specific challenge: the label “Not mentioned” was often used either as a legitimate answer or as a fallback after other options were applied. Preprocessing had to account for this dual role, retaining “Not mentioned” only when it was the sole response across all slots, and otherwise aggregating all informative answers into a deduplicated, comma-separated string.

To standardize the dataset, answer text was cleaned to remove extraneous tokens (i.e., brackets, quotation marks, and “(please specify)” suffixes). Prompt formatting involved synthesizing a structured input per image, which included:

1. Cleaned question text (with numeric prefixes and instructional phrases removed),
2. Question type and category metadata,
3. Synthesized clinical background from query title and content,
4. Comma-separated list of possible answer choices.

Each sample was wrapped in a constrained conversational prompt format, explicitly instructing models to return only the exact answer text(s)—comma-separated if multiple labels were valid.

Image paths were validated using PIL to ensure all images could be opened; corrupted or missing files were excluded. The final preprocessed data was serialized into batched `.pk1` files (100 samples per batch), each containing structured dictionaries with the fields: `encounter_id`, `base_qid`, `query_text`, `image_path`, `answer_text`, `question_type`, `question_category`, and a multi-label indicator.

5. Methodology

We evaluated six architectural configurations for medical visual question answering, systematically comparing base and fine-tuned vision-language models both independently and as components within

enhanced reasoning frameworks. Each configuration was tested across seven open-source models: LLaMA-3.2-11B-Vision, Qwen2-VL (2B, 7B), Qwen2.5-VL (3B, 7B), and Gemma-3 (4B, 12B). For agentic reasoning and retrieval-augmented generation (RAG), we additionally employed Gemini 2.5 Flash as the instruction-following model responsible for multi-step reasoning and aggregation.

1. Base models performing direct inference on medical VQA tasks
2. Fine-tuned models (LoRA-adapted on ImageCLEFmedical 2025 data) performing direct inference
3. Reasoning layer enhancement utilizing base model inference
4. Reasoning layer enhancement utilizing fine-tuned model inference
5. Agentic RAG system utilizing base model inference
6. Agentic RAG system utilizing fine-tuned model inference

5.1. Model Fine-Tuning

Several pretrained, open-source vision-language models were fine-tuned on the processed MEDIQA-MAGIC dataset for diagnostic question answering. The models included LLaMA-3.2-11B-Vision-Instruct, Gemma-3 (4B and 12B), and Qwen2/2.5-VL (2B, 3B, and 7B variants). This range enabled comparison of performance across architectures and scales and, later, the exploration of ensemble-like strategies that reflected clinical workflows involving multiple expert perspectives.

Model-specific preprocessing during training included chat template application with appropriate special token handling for different architectures (LLaMA, Qwen), RGB image format standardization, and label masking for special tokens to ensure proper loss computation. While we initially tested modifying chat templates to jointly process all images in a single forward pass, this approach led to out-of-memory (OOM) errors. Despite exploring various optimizations, the backward pass failed due to the need to retain gradients across all image-conditioned operations. As a result, we proceeded by passing in one image at a time for open-source training and inference.

To reduce memory usage and training time, parameter-efficient fine-tuning was applied using Low-Rank Adaptation (LoRA) with 4-bit quantization via BitsAndBytes. LoRA was configured with rank 8, alpha 16, and dropout of 0.05, applied to attention projection layers (specifically `q_proj` and `v_proj` for LLaMA and Qwen models, or all linear layers for other architectures). For non-LLaMA models, the language modeling head and embedding layers were additionally included in trainable parameters. All models utilized NF4 quantization with double quantization enabled, using `bf16` as the compute dtype.

Each training sample combined the preprocessed prompt with its associated medical image, processed through model-specific vision encoders. The training employed gradient accumulation over 32 steps with a per-device batch size of 1, effectively achieving a batch size of 32. Models were trained for 3 epochs using the fused AdamW optimizer with a learning rate of $1e-4$, gradient clipping at 0.3, and a constant learning rate schedule with 3% warmup ratio. Gradient checkpointing with non-reentrant mode was enabled to further optimize memory usage.

Training was conducted on NVIDIA A100 80GB GPUs within Georgia Tech's high-performance computing platform, the Partnership for an Advanced Computing Environment (PACE). Each model required approximately 10 hours to complete, totaling 70 GPU-hours across all configurations. The Supervised Fine-Tuning Trainer (SFTTrainer) from the TRL library was used, with checkpoints saved every 50 steps and training progress monitored via TensorBoard. After training, LoRA adapters were merged into the base models to create standalone versions for inference, with all artifacts saved using safe serialization and 2GB maximum shard size.

5.2. Model Inference

The inference pipeline was designed to handle both base and fine-tuned models, with automatic selection based on configuration parameters. For inference on finetuned models, models were loaded with the same quantization settings used during training (4-bit NF4 quantization with `bf16` compute dtype) to maintain consistency and reduce memory requirements. The inference process utilized a specialized `MedicalImageInference` class that handled model loading, input preprocessing, and prediction generation. For each test sample, the system constructed a standardized prompt using the same template as training, combining the clinical context, question text, and available answer options with the corresponding medical image. The prompt explicitly tested and instructed the model to respond only with the exact text of applicable options without explanations, handle multi-label cases with comma separation, and default to "Not mentioned" when uncertain.

During inference, generation parameters were carefully tuned to balance output quality and diversity, using temperature 0.9, top-p 0.95, top-k 64, and a maximum of 100 new tokens with sampling enabled. The system processed predictions in batches through preprocessed pickle files, automatically handling different batch prefixes for validation and test datasets. Post-processing steps included removing any system artifacts, special tokens, or formatting prefixes that might appear in the generated text. For multi-answer questions, predictions from multiple images of the same encounter were aggregated using an aggregation mechanism that respected question-specific maximum answer limits (ranging from 1 to 9 depending on the question type). The consolidation process counted prediction frequencies across images, selected the most common responses up to the allowed limit, and handled ties through deterministic random selection with a fixed seed. Final predictions were formatted according to the official evaluation requirements, mapping text answers to their corresponding indices and distributing multiple answers across question variants when applicable. The complete pipeline generated both CSV files with prediction details and JSON files formatted for official submission, along with empty mask prediction directories as required by the competition format.

5.3. Reasoning Layer

The reasoning layer acted as a senior dermatologist reviewing multiple opinions and evidence to reach a final diagnosis. Unlike traditional ensemble methods that rely on majority voting or weighted averaging of predictions, this system performed interpretive synthesis that mirrored how senior clinicians integrate diverse expert opinions during complex case reviews. Where conventional ensembles might simply count votes, our reasoning layer evaluated the quality of evidence, considered clinical context, and applied domain-specific knowledge to reach conclusions—sometimes overriding majority predictions when evidence warranted it.

We designed a multi-stage process using `gemini-2.5-flash-preview` [16] that systematically enriched raw inputs before making final decisions:

Stage 1: Image Analysis and Aggregation. The model extracted standardized dermatological features from each image, including lesion morphology (flat, raised, depressed), precise anatomical locations, color characteristics, texture patterns, and distribution. For encounters with multiple images, these individual analyses were synthesized into a unified assessment that captured the complete clinical picture—identifying patterns across images while preserving important variations.

Stage 2: Clinical Context Extraction. The system processed accompanying clinical text to extract structured information including patient demographics, symptom duration and progression, identified

triggers, and relevant medical history. This preprocessing transformed free-text clinical notes into a consistent JSON format, enabling more reliable integration with visual findings.

Stage 3: Evidence-Based Reasoning. In the final stage, the model synthesized these enriched analyses with predictions from other models through carefully engineered dynamic and query-specific prompts. Other models' predictions were explicitly framed as "advisory inputs that may contain errors" rather than ground truth, preventing the system from defaulting to simple majority agreement. The reasoning layer was instructed to critically evaluate all evidence and provide step-by-step justification for its conclusions.

We also evaluated Gemini 2.5 Flash independently, prompting it directly with image and context information. While its standalone performance showed promise, it was notably weaker than the results obtained using our multi-stage reasoning setup that incorporated model predictions and structured synthesis. This highlighted the benefit of combining Gemini's instruction-following capabilities with intermediate expert signals and multimodal preprocessing to better reflect clinical decision-making.

This approach addressed key limitations of traditional ensembles. For instance, when multiple models predicted "size of thumb nail" for a case showing numerous small lesions distributed across a palm-sized area, the reasoning layer correctly identified this as "size of palm" based on the aggregate affected area rather than individual lesion size. The system employed specialized reasoning strategies for different question types: size assessments relied exclusively on visual evidence from the image analysis, deliberately excluding potentially misleading clinical descriptions, while color evaluations distinguished between uniform presentations and multi-tonal patterns that would indicate a "combination" answer.

By implementing clinical consultation practices in AI systems, this reasoning layer ensured that final diagnoses reflected not just statistical consensus but genuine medical reasoning—evaluating evidence quality, reconciling contradictory inputs, and applying contextual knowledge to reach accurate conclusions.

5.4. Agentic Retrieval-Augmented Generation

We developed a multi-agent retrieval-augmented generation system using Gemini 2.5 Flash [16], designed to emulate how dermatologists combine visual assessment with targeted reference consultation. Rather than following a fixed sequence, the system distributes tasks across specialized agents for evidence integration, reasoning, reflection, and reanalysis—each adapting its behavior based on the available evidence, diagnostic ambiguity, and question complexity.

5.4.1. Input Layer

The input layer collects all core elements of the clinical encounter, including patient-provided images along with a description of their symptoms and concerns. We also passed diagnostic predictions from several large vision-language models from our earlier steps. A curated knowledge base is included as an additional source of external reference.

5.4.2. Context Assembly

The context assembly layer consists of five specialized agents operate in a modular sequence, each performing a distinct function, image analysis, clinical context extraction, evidence integration, diagnostic reasoning, and iterative refinement, collectively enabling adaptive, multi-stage decision-making.

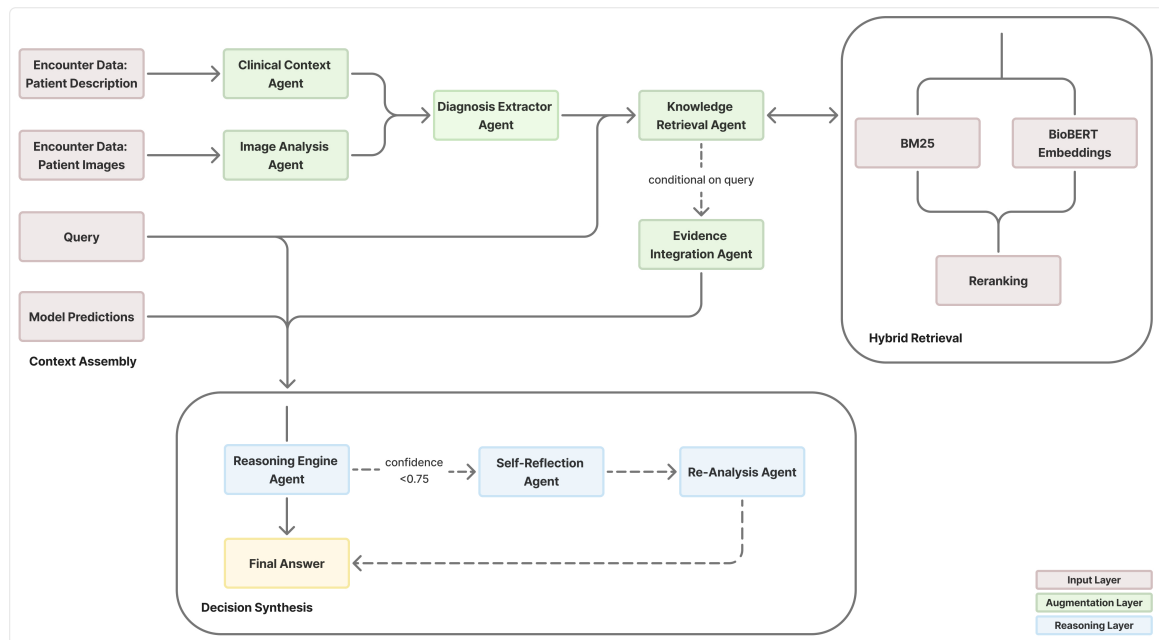


Figure 2: Agentic RAG architecture featuring a multi-stage pipeline with integrated agents that process queries, encounter data, and model predictions through collaborative reasoning and iterative refinement..

Image Analysis Agent The Image Analysis Agent extracts structured features from dermatological images, including lesion size, location, shape, color, and distribution. By analyzing multiple images per encounter, it builds a more complete view of the skin condition. The agent also identifies signs of scratching or trauma and visual indicators that may suggest duration, such as healing or chronic changes, to support more informed downstream reasoning.

Clinical Context Agent The Clinical Context Agent extracts structured, medically relevant details from the patient's written description, organizing them into consistent categories such as demographics, lesion location, appearance, symptom duration, and prior history. It also identifies mentions of itching, pain, and potential triggers. Because the input text is patient-authored, the agent implicitly filters out noise—like emojis or repetition—to provide a cleaner, medically relevant summary that supports more accurate downstream analysis.

Diagnosis Extractor and Knowledge Retrieval Agents with Hybrid Retrieval The Diagnosis Extractor combines visual findings from Image Analysis with structured clinical input from the Clinical Context Agent to generate diagnostic hypotheses grounded in both modalities. These hypotheses guide the Knowledge Retrieval Agent, which formulates targeted search queries based on the suspected conditions, integrated multimodal context, and question information. Rather than issuing generic queries, the agent dynamically composes prompts that reflect the likely diagnosis and clinical focus, such as location, symptoms, or morphology.

For example, given a question like "Where is the affected area?" and extracted diagnoses like eczema and dermatitis, the agent could generate queries such as "eczema common body locations" or "dermatitis site distribution patterns," tailoring search prompts to both the diagnosis and context of the question. In this way, the system actively guides knowledge retrieval using structured hypotheses, rather than

simply retrieving information in response to the original input.

These tailored queries form the input to a hybrid search strategy that combines BM25 keyword matching with semantic similarity search based on dense embeddings. Semantic search is performed using the pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb model, a specialized variant of BioBERT fine-tuned on a diverse set of natural language inference and semantic similarity benchmarks including SNLI, MNLI, SCINLI, SCITAIL, MEDNLI, and STS-B [17]. Built on biomedical domain pretraining, this model produces 768-dimensional embeddings that support nuanced generalization across clinical and scientific text. These embeddings enable retrieval of passages from LanceDB that are semantically aligned with the query, even when there is minimal lexical overlap.

In parallel, BM25 keyword search is performed using BM25Okapi, which scores documents based on term frequency (TF), inverse document frequency (IDF), and document length normalization. This method prioritizes documents containing exact or rare keyword matches, particularly useful for surfacing medically significant terms such as drug names, anatomical locations, or uncommon conditions that may not be captured by semantic models. The results from both semantic and keyword searches are concatenated and deduplicated using document IDs to avoid redundancy.

To improve retrieval precision, the system applies the cross-encoder/ms-marco-MiniLM-L6-v2 model to rerank the combined candidate passages. This cross-encoder, trained on the MS MARCO passage ranking task, jointly encodes each query-document pair to capture fine-grained relevance. It achieves a strong MRR@10 of 39.01 on the MS MARCO development set, outperforming many larger models while maintaining efficiency with only 22.7 million parameters [18]. Unlike bi-encoders that embed queries and documents independently, the cross-encoder models their full interaction, making it particularly effective at capturing subtle clinical nuances where phrasing differences may alter meaning. This step is especially critical in medical contexts, where ambiguous matches can lead to misinformation or retrieval failures. The final top-k results are selected based on the highest cross-encoder relevance scores.

The knowledge base supporting this pipeline is constructed from the Hugging Face dataset brucewayne0459/Skin_diseases_and_care, which contains over 800 curated entries sourced from the American Academy of Dermatology Association. These entries span dermatologic conditions, treatments, prevention guidelines, and skin, hair, and nail care. All documents are embedded and stored in LanceDB, a high-performance vector database optimized for fast semantic retrieval at scale, with support for hybrid and filtered search operations.

Finally, retrieval behavior is conditioned on the type of question being asked. For example, RAG is disabled for image-dependent questions such as those concerning lesion color or size, where direct visual analysis provides more accurate answers. This diagnosis-first, context-aware retrieval pipeline ensures that external knowledge is only surfaced when clinically appropriate, thereby improving both the relevance and specificity of information used by downstream agents. Each stage of the pipeline, from diagnosis extraction to reranking, is modular and designed for clinical robustness.

Evidence Integration Agent The Evidence Integration Agent synthesizes outputs from three modalities, image analysis, structured clinical context, and, when available, retrieved medical knowledge, into a unified representation tailored to the question type. It applies adaptive, task-specific weights to each source: visual cues are emphasized for appearance-based questions (i.e., lesion color), clinical history takes precedence for treatment or symptom-related prompts, and retrieved knowledge contributes more heavily to diagnosis-driven tasks. The agent constructs a structured prompt embedding these inputs and weights, which is then passed to Gemini for reasoning. The model returns a JSON-formatted output capturing integrated findings, source concordance or contradiction, and a weighted summary of key features. This module provides a context-aware synthesis that downstream agents can build upon for

more accurate and interpretable decision-making.

5.4.3. Decision Synthesis

Three interconnected agents, Reasoning, Self-Reflection, and Re-Analysis, form the system’s decision-making backbone, working in sequence to generate, evaluate, and refine diagnostic predictions based on the integrated evidence.

Reasoning Engine Agent The Reasoning Engine generates a diagnostic prediction using the question text and type, answer options, integrated evidence with pre-assigned weights, and pre-existing model predictions sourced from multiple LLMs—both fine-tuned and general-purpose. These model outputs are treated as contextual signals rather than authoritative inputs; the agent is prompted to avoid defaulting to consensus and must explicitly justify any alignment with model suggestions. The agent selects an answer, assigns a confidence score, and produces a structured explanation grounded in the weighted evidence. This output serves as a provisional decision that can later be scrutinized or revised by downstream agents.

Self-Reflection Agent The Self-Reflection Agent introduces metacognitive oversight by reassessing the Reasoning Engine’s output when the confidence score falls below a predefined threshold (i.e., 0.75), signaling uncertainty. It revisits the same inputs, questions, evidence, model predictions, and evaluates the initial answer for overlooked or misinterpreted evidence, reasoning gaps, and the appropriateness of the assigned confidence. Rather than relying on fixed rules, the agent makes a qualitative determination about whether a revision is warranted. If so, it sets the `requires_revision` flag to `True` and prepares the case for deeper reanalysis; if not, it affirms the original reasoning, possibly with an adjusted confidence score. This reflection layer strengthens diagnostic accountability by allowing the system to identify and flag potential misjudgments before proceeding.

Re-Analysis Agent The Re-Analysis Agent engages only when the Self-Reflection Agent flags the need for revision, conducting a deeper reassessment of the original question, integrated evidence, initial reasoning, and critique. Rather than reiterating earlier outputs, it systematically reconsiders the case with attention to previously identified flaws—such as overlooked evidence or misweighted signals—and generates a revised or reaffirmed answer, updated confidence score, and structured rationale. This final step allows the system to recover from low-confidence or error-prone decisions, reinforcing diagnostic precision while modeling a closed-loop process of autonomous reasoning and correction.

By incorporating this third layer, the pipeline models a closed-loop cognitive system: it makes an initial prediction, introspects for quality, and, when needed, revisits and improves upon its own reasoning. This architecture enables the system to exhibit autonomous judgment, adaptive error correction, and explainable decision-making—capabilities that are especially valuable in high-stakes clinical domains like dermatology. This system introduces several innovations that extend beyond traditional RAG pipelines, enabling more robust, flexible, and clinically grounded diagnostic reasoning. By combining specialized agents, autonomous feedback loops, and adaptive evidence handling, the architecture mimics key aspects of expert decision-making in dermatology.

6. Results

To evaluate model performance across the CVQA task, accuracy was compared across seven vision-language models as well as two reasoning configurations using baseline and finetuned models separately.

Baseline models achieved moderate accuracy, and fine-tuning had mixed effects across different architectures. Fine-tuning decreased overall accuracy for five of the seven base models (Qwen2-2b, Qwen2-7b, Qwen2.5-7b, Gemma3-4B, and LLaMA-3.2-11B), with validation accuracy drops ranging from 8% (LLaMA-3.2-11B) to as high as 29% (Qwen2.5-7B), implying varying degrees of overfitting. Only two models, Qwen2.5-3B and Gemma3-12B, showed improvement after fine-tuning, gaining roughly 2% and 11% in accuracy respectively. These inconsistent results can be observed in Table 4 and Figure 4 (see Appendix), which contrast each model’s average validation accuracy before and after fine-tuning. Qwen2.5-7B without finetuning achieved the highest accuracy of the base and finetuned models, whereas Qwen2.5-3B had the highest accuracy among fine-tuned models. These results indicate that fine-tuning effects varied substantially by model.

Contrasting the individual models, the combined-model methods, including the reasoning layer and agentic RAG, yielded the highest overall accuracies regardless of whether they utilized base or finetuned model predictions as input. This is highlighted in Figure 1, which shows that advanced models with integrative reasoning and retrieval performed better than the best performing base model across most of the questions on the validation dataset except for select questions, CQID020, CQID020, and CQID034. The difference between the different architectures when comparing base models is fairly comparable. On the other hand, the best performing fine-tuned model achieved meaningfully worse accuracy on the validation dataset when compared to the combined-model methods. The only question that showed marked improvement was Site (CQID010).

Table 1 and Table 2 further demonstrate how the reasoning layer (which aggregated multiple model predictions) achieved the best average performance on both the validation and test sets. For instance, the reasoning layer achieved 71.2% accuracy on the validation data, slightly higher than the agentic RAG’s 69.0% and the best single model’s 67.2%. The gap was more pronounced on the test dataset where the best individual non-finetuned model (Qwen2.5-7B) dropped to 37.5% accuracy while the reasoning layer and agentic RAG retained 70.6% and 69.2% accuracy respectively. While all three methods performed similarly on the validation dataset, the combined-model methods maintained consistent, nearly double the performance of the best single model on the test dataset.

The submitted test results shown below reflect the official outputs we submitted to the MEDIQA-MAGIC 2025 leaderboard. These were produced using our final reasoning and RAG systems, with no further test-time tuning.

Table 1

Base model performance comparison across validation and test datasets with three architectures: a) best performing model inference, b) Reasoning Layer using base model predictions, and c) Agentic RAG using base model predictions. The Reasoning Layer achieves the highest average performance on both datasets.

Question	Validation Dataset			Test Dataset (Submitted Results)		
	Qwen2.5-VL-7B	Reasoning Layer	Agentic RAG	Qwen2.5-VL-7B	Reasoning Layer	Agentic RAG
CQID010	0.4821	0.5714	0.5357	0.3100	0.5100	0.4700
CQID011	0.8333	0.9048	0.8762	0.3847	0.8403	0.8552
CQID012	0.6086	0.7083	0.7009	0.5317	0.6967	0.6900
CQID015	0.7679	0.8929	0.8571	0.3100	0.8500	0.8500
CQID020	0.5708	0.5653	0.5771	0.3122	0.5587	0.5561
CQID025	0.8929	0.8036	0.8036	0.4200	0.8700	0.8400
CQID034	0.4643	0.4286	0.3929	0.0100	0.5500	0.5100
CQID035	0.8750	0.8929	0.8214	0.7200	0.8100	0.8200
CQID036	0.5536	0.6429	0.6429	0.3700	0.6700	0.6400
Average	0.6721	0.7123	0.6898	0.3743	0.7062	0.6924

Table 2

Fine-tuned model performance comparison across validation and test datasets with two architectures: a) Reasoning Layer using finetuned model predictions, and b) Agentic RAG using finetuned model predictions. Qwen2.5-VL-7B was excluded from inference on test due to overfitting during training. The Reasoning Layer maintains the highest average performance on both datasets.

Question	Validation Dataset		Test Dataset (Submitted Results)	
	Reasoning Layer	Agentic RAG	Reasoning Layer	Agentic RAG
CQID010	0.6071	0.5536	0.5300	0.4400
CQID011	0.8777	0.8795	0.8683	0.8363
CQID012	0.6815	0.7173	0.6625	0.6858
CQID015	0.8214	0.8214	0.8100	0.7800
CQID020	0.5821	0.5421	0.5649	0.5544
CQID025	0.8214	0.8036	0.8900	0.8600
CQID034	0.4643	0.3929	0.6000	0.4800
CQID035	0.8929	0.8393	0.8100	0.7900
CQID036	0.6071	0.6071	0.6500	0.6500
Average	0.7062	0.6841	0.7095	0.6752

7. Discussion

The divergent impact of fine-tuning on different models suggests that model architecture and pre-training state heavily influence how a model benefits from additional training data. Our results showed that some larger and strong-performing models (i.e. Qwen2.5-7B) lost validation accuracy after domain-specific fine-tuning, whereas others (i.e. Gemma3-12B) gained considerable accuracy (11%) from the same fine-tuning methods. One possible explanation is that models with high baseline performance

on the CVQA task may have already aligned well with the task distribution; fine-tuning them on a limited training set could induce overfitting or disrupt previously learned generalizable features, a form of catastrophic forgetting [19]. In contrast, models with lower initial accuracy had more room for improvement and likely benefited from learning domain-specific patterns present in the fine-tuning data. Sensitivity to fine-tuning has been observed in other multi-modal medical models as well. Med-Gemini achieved substantial performance gains through targeted fine-tuning on medical data [5]. These findings nuance previous results by showing such gains are not universal. The efficacy of fine-tuning may depend on the model's architecture and how well its pre-training corpus covered dermatological concepts. To improve fine-tuning outcomes, adaptive strategies may be required to account for each model's idiosyncrasies, especially for highly capable models where fine-tuning may hurt performance. These results highlight the need for careful model-specific tuning approaches and validation as a "one-size-fits-all" fine-tuning can yield inconsistent outcomes in multi-modal medical AI systems.

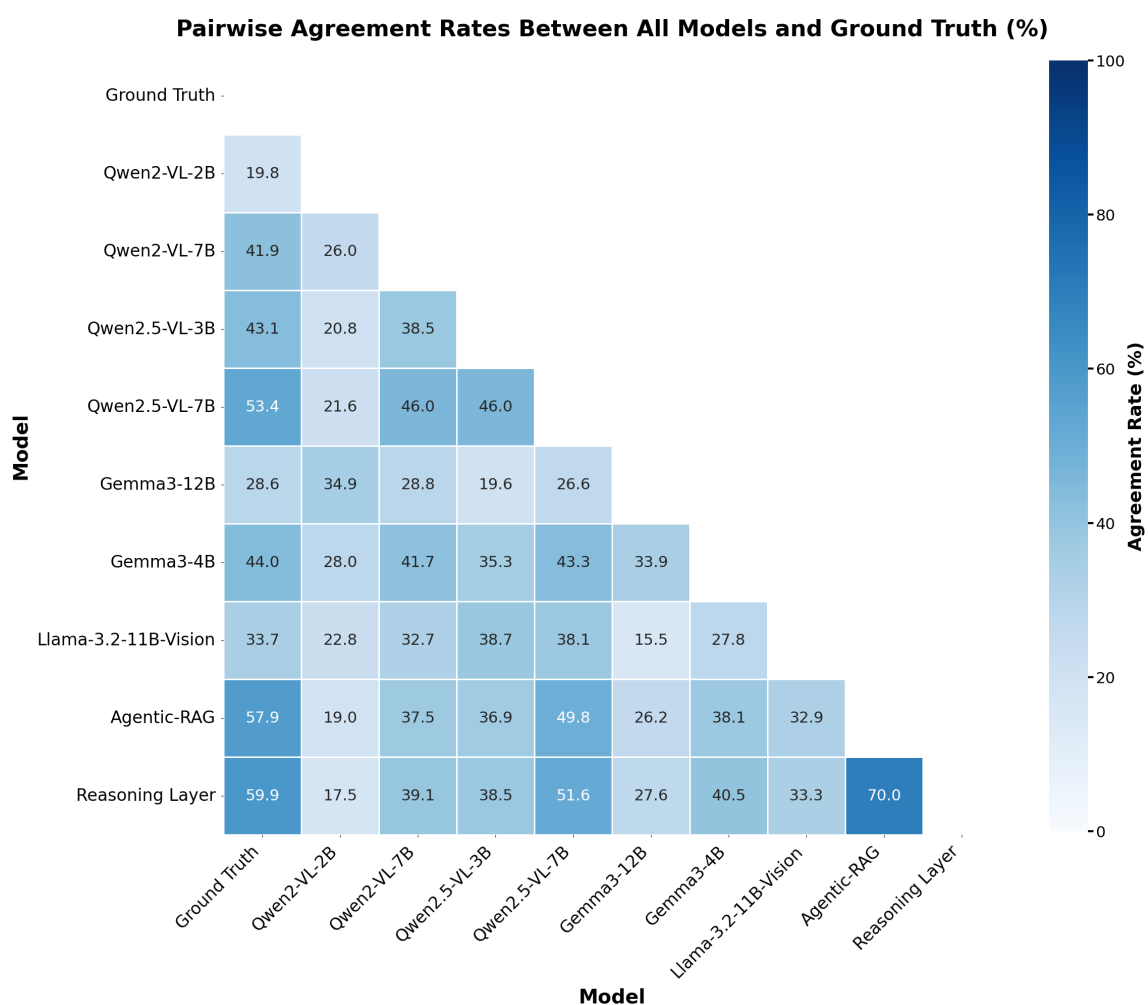


Figure 3: Pairwise agreement rates (%) among base model predictions and ground truth labels. Higher values indicate greater consistency in labels.

In contrast to individual models, the reasoning layer, which aggregated multiple model predictions,

delivered a consistent pronounced performance boost, especially on unseen test cases. This improvement is analogous to an ensemble-like diagnosis in clinical practice where multiple independent opinions are combined to reach a more reliable conclusion. Figure 3 demonstrates the pairwise agreements among the models and ground truth labels. Here we see significant model diversity with little agreement ranging in values 17.5% to 51.6%, a key trait required for effective ensemble-like methods. By utilizing multiple outputs from different models and taking the clinical context into mind, the reasoning layer likely diversified the errors and captured strengths of each model. The net effect was a more robust system less prone to a single model's failure as exemplified in Table 6 (see Appendix). This aggregated reasoning approach aligns with previous research where an iterative consensus ensemble of large language models achieved up to a 27% increase in QA accuracy compared to any single model [20]. These results confirm that a similar principle holds in multi-modal CVQA: an ensemble simulating a "committee" of AI diagnosticians can produce more accurate and reliable answers. This is promising clinically as it mirrors the workflow of clinicians where difficult cases are often discussed to reduce error. The performance gain observed from the reasoning layer underscores the value of incorporating multi-source reasoning into diagnostic AI systems, especially for safety-critical applications where erroneous single-model outputs could lead to clinically significant misdiagnoses, which is a major concern for AI diagnostic tools.

The agentic retrieval-augmented generation module introduced additional complexity allowing the system to dynamically fetch and integrate external medical knowledge during question answering. Quantitatively, the agentic RAG model performed similarly to the reasoning layer (within 3% accuracy) and 70% pairwise agreement per Figure 3, indicating that retrieval alone did not dramatically exceed the reasoning layer's already strong performance. However, focusing only on accuracy metrics would overlook the important qualitative contributions of the agentic RAG. By retrieving relevant dermatology references (such as disease descriptions and treatment guidelines) and weaving them into the answer rationale, the agentic RAG system produced responses that were often richer in context and explanation. For instance, when faced with an ambiguous lesion description, the RAG component could pull in a textbook snippet about similar presentations, thereby providing a fuller and grounded justification for the chosen diagnoses.

An example of this is shown in Table 7. Even if this extra knowledge did not always encourage the model to select an answer consistent with ground truth labels, it adds interpretative and explainability value grounded in clinical context that individual models lack. The use of an agentic RAG system bridges the gap between the limited patient-provided context and the broader medical knowledge base. As a result the system's answer better resembled an informed clinical explanation rather than a narrow pattern match. This explanation is arguably more valuable for clinicians and professionals that work with AI technologies and want rationale behind answers from AI. It is easier for professionals to detect hallucinations and the trustworthiness of an AI system whose answers are grounded in medical expertise by citing its sources. Our methodology explicitly drew inspiration from how human clinicians consult external resources. The agentic RAG module's value may not fully manifest in blunt accuracy percentages, but it enhances the trustworthiness and depth of responses. For telemedicine applications, explanations are vital for patients and providers who want to know why a certain diagnosis was reached.

8. Future Work

While our system showed strong performance in both diagnostic accuracy and clinical plausibility, there are several areas that could be improved to support real-world use and deeper scientific understanding. For instance, inference efficiency is a key concern for clinical deployment. The reasoning layer produced responses in roughly one minute per query, whereas the agentic RAG system—due to its multi-step prompt chaining—took approximately seven minutes. This latency presents a barrier to real-time use. Future work should explore optimizations such as prompt compression, parallelized reasoning, and response caching to reduce inference time without compromising reasoning quality. The retrieval component is another area with room for enhancement. Our system used a fixed corpus, which may have limited the relevance and diversity of supporting evidence. Expanding the knowledge base to include additional medical sources, such as dermatology reference texts, structured ontologies, or clinical guidelines, could help improve the specificity and clinical depth of generated answers.

We also plan to revisit our fine-tuning strategy, as several large models showed reduced performance after training. This points to the need for more stable approaches, such as adaptive regularization, early stopping based on clinically meaningful metrics, or curriculum-based training. In parallel, we aim to apply this framework to classification-style VQA datasets to evaluate how well the system generalizes across different task formats. Another area for refinement is the agentic RAG system’s reasoning flow. We observed some redundancy in its multi-step prompts, with repeated logic across stages. While this may have added robustness, it also introduced unnecessary complexity. Future iterations should focus on streamlining these reasoning chains to improve efficiency without sacrificing interpretability.

Beyond system improvements, there is a growing consensus that evaluation of medical AI should extend beyond traditional metrics to encompass the quality of reasoning, justification and the decision-making value provided to human clinicians [21]. An AI system’s utility in health is not just about how often it selects the correct label, but also whether its reasoning process is sound and its advice can be trusted in practice. In a tele-medical scenario, additional factors like how the AI system communicates a differential diagnoses or its ability to incorporate new patient information on the fly are crucial for adoption. For future work we advocate for more nuanced evaluation frameworks where reasoning quality and clinical usefulness are measured in addition to accuracy.

9. Conclusion

This work highlights the potential of combining fine-tuned vision-language models with reasoning and agentic retrieval-augmented generation (RAG) to support dermatological diagnosis. While fine-tuning individual models produced variable results, integrating a reasoning layer and agentic RAG led to stronger performance by synthesizing outputs across models and anchoring predictions in trusted dermatological knowledge.

The dataset revealed important real-world challenges: some patients declined pathological exams due to financial constraints, while others had already sought care unsuccessfully at multiple hospitals. These cases reflect a broader healthcare access gap, patients in need of accurate diagnosis who cannot obtain it through traditional means, and where AI-assisted telemedicine presents a promising alternative.

Yet a major barrier to deploying AI in clinical practice remains: the opacity of large language models and deep learning systems. These models often produce predictions through mechanisms that are not easily interpretable by clinicians, limiting their utility in high-stakes medical contexts where trust and

transparency are essential.

Our proposed methods, namely agentic rag, directly sought to address this challenge. Instead of offering opaque outputs, it provides traceable, context-aware justifications grounded in clinical literature and guidelines. This shifts AI from a black-box predictor to a collaborative decision support tool, one that clinicians can inspect, understand, and ultimately rely on.

Importantly, our results show that interpretability does not require sacrificing accuracy. By delivering both high diagnostic performance and transparent reasoning, our system illustrates a viable path forward for trustworthy AI in healthcare. We believe this framework, reasoning-enhanced, knowledge-grounded, and agentic, can extend beyond dermatology to other domains of telemedicine, particularly where patients face similar structural barriers to care. This work is a step toward AI systems that support equitable, auditable, and expert-aligned clinical decision-making.

Acknowledgments

We thank the Georgia Tech Applied Research Competitions team for their invaluable support throughout this research. This research was supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment at the Georgia Institute of Technology, Atlanta, Georgia, USA.

Thanks to the developers of ACM consolidated LaTeX styles <https://github.com/borisveytsman/acmart> and to the developers of Elsevier updated L^AT_EX templates <https://www.ctan.org/tex-archive/macros/latex/contrib/els-cas-templates>.

Declaration on Generative AI

The authors have not employed any Generative AI tools in writing.

References

- [1] W. Yim, A. Ben Abacha, N. Codella, R. A. Novoa, J. Malvey, Overview of the mediqa-magic task at imageclef 2025: Multimodal and generative telemedicine in dermatology, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
- [2] W. Yim, Y. Fu, A. Ben Abacha, M. Yetisgen, N. Codella, R. A. Novoa, J. Malvey, Dermavqa-das: Dermatology assessment schema (das) and datasets for closed-ended question answering and segmentation in patient-generated dermatology images, CoRR (2025).
- [3] E. Okon, V. Rachakonda, H. J. Hong, C. Callison-Burch, J. B. Lipoff, Natural language processing of reddit data to evaluate dermatology patient experiences and therapeutics, Journal of the American Academy of Dermatology 83 (2020) 803–808. doi:10.1016/j.jaad.2019.07.014.
- [4] J. Zhou, X. He, L. Sun, J. Xu, X. Chen, Y. Chu, L. Zhou, X. Liao, B. Zhang, X. Gao, Skingpt-4: An interactive dermatology diagnostic system with visual large language model, arXiv preprint arXiv:2304.10691 (2023). doi:10.48550/arXiv.2304.10691.
- [5] L. Yang, S. Xu, A. Sellergren, et al., Advancing multimodal medical capabilities of gemini, arXiv preprint arXiv:2405.03162 (2024). doi:10.48550/arXiv.2405.03162.

- [6] J. Liu, Y. Wang, J. Du, J. T. Zhou, Z. Liu, Medcot: Medical chain of thought via hierarchical expert, arXiv preprint arXiv:2412.13736 (2024). doi:10.48550/arXiv.2412.13736, presented at EMNLP 2024.
- [7] X. Hu, J. Wang, J. Hamm, R. R. Yotsu, Z. Ding, Enhancing skin disease diagnosis: Interpretable visual concept discovery with sam, arXiv preprint arXiv:2409.09520 (2024). doi:10.48550/arXiv.2409.09520, accepted at WACV 2025.
- [8] X. Chi, R. Zhang, Z. Jiang, et al., Mchat: Empowering vlm for multimodal llm interleaved text-image generation, arXiv preprint arXiv:2311.17963 (2023). doi:10.48550/arXiv.2311.17963.
- [9] W. Huang, A. Wu, Y. Yang, et al., Llm2clip: Powerful language model unlocks richer visual representation, arXiv preprint arXiv:2411.04997 (2024). doi:10.48550/arXiv.2411.04997.
- [10] P. R. Bassi, Q. Wu, W. Li, et al., Label critic: Design data before models, arXiv preprint arXiv:2411.02753 (2024). doi:10.48550/arXiv.2411.02753.
- [11] G. Xiong, Q. Jin, Z. Lu, A. Zhang, Benchmarking retrieval-augmented generation for medicine, arXiv preprint arXiv:2402.13178 (2024). doi:10.48550/arXiv.2402.13178.
- [12] J. Sohn, Y. Park, C. Yoon, et al., Rationale-guided retrieval augmented generation for medical question answering, arXiv preprint arXiv:2411.00300 (2024). doi:10.48550/arXiv.2411.00300.
- [13] M. H. Vu, T. Löfstedt, T. Nyholm, R. Sznitman, A question-centric model for visual question answering in medical imaging, IEEE Transactions on Medical Imaging (2020). doi:10.1109/TMI.2020.2978284.
- [14] S. Tascon-Morales, P. Márquez-Neila, R. Sznitman, Consistency-preserving visual question answering in medical imaging, arXiv preprint arXiv:2206.13296 (2022). URL: <https://arxiv.org/abs/2206.13296>. doi:10.48550/arXiv.2206.13296.
- [15] J. Parekh, P. Khayatan, M. Shukor, et al., A concept-based explainability framework for large multimodal models, arXiv preprint arXiv:2406.08074 (2024). doi:10.48550/arXiv.2406.08074, accepted at NeurIPS 2024.
- [16] G. DeepMind, Gemini, <https://deepmind.google/technologies/gemini/>, 2024. Accessed: 2025-05-26.
- [17] P. Deka, A. Jurek-Loughrey, et al., Evidence extraction to validate medical claims in fake news detection, in: International Conference on Health Information Science, Springer, 2022, pp. 3–15.
- [18] N. Reimers, I. Gurevych, Cross-encoder for ms marco: Minilm-l6-v2, <https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2>, 2021. Accessed: 2025-05-25.
- [19] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, Y. Zhang, An empirical study of catastrophic forgetting in large language models during continual fine-tuning, 2025. URL: <https://arxiv.org/abs/2308.08747>. arXiv:2308.08747.
- [20] M. Omar, B. S. Glicksberg, G. N. Nadkarni, E. Klang, Refining llms outputs with iterative consensus ensemble (ice), medRxiv (2024). URL: <https://www.medrxiv.org/content/early/2024/12/27/2024.12.25.24319629>. doi:10.1101/2024.12.25.24319629.
- [21] I. D. Raji, R. Daneshjou, E. Alsentzer, It’s time to bench the medical exam benchmark, NEJM AI 2 (2025) AIE2401235. URL: <https://ai.nejm.org/doi/full/10.1056/AIE2401235>. doi:10.1056/AIE2401235. arXiv:<https://ai.nejm.org/doi/pdf/10.1056/AIE2401235>.

A. Appendix

Table 3: Expected answers for Encounter 6 reveal key mismatches with ground truth: “size” included contradictory labels despite widespread lesions, “color” listed a combination when only pink was visible, and “texture” was marked “Not mentioned” despite clearly rough, raised features. These inconsistencies may have penalized more clinically accurate model predictions.

Clinical Case Analysis	
Case Information and Images	Ground Truth Labels
Encounter: ENC00006 Query Title: "Can everyone diagnose what this skin disease is?" Query Content: "The patient is a 50-year-old male construction worker who is frequently exposed to cement hardener. The first image is of the chest, and the second image is of the back. The patient has hand rashes with peeling and cracking, followed by the appearance of rashes on the chest and back, which are unbearably itchy." Images: 	CQID010: How much of the body is affected? "Widespread" CQID011: Where is the affected area? "Chest/abdomen, back, upper extremities" CQID012: How large are the affected areas? "Size of palm, larger area" CQID015: When did the patient first notice? "Not mentioned" CQID020: What label best describes? "Crust, raised or bumpy, scab" CQID025: Associated itching? "Yes" CQID034: Color of skin lesion? "Combination (please specify)" CQID035: How many skin lesions? "Multiple (please specify)" CQID036: Skin lesion texture? "Not mentioned"

Table 4

Base model performance comparison across language models on validation dataset. Qwen2.5-VL-7B achieves the highest average performance.

Question	Qwen2-VL-2B	Qwen2-VL-7b	Qwen2.5-VL-3B	Qwen2.5-VL-7B	Gemma3-4B	Gemma3-12B	LLaMA-3.2-11B-Vision
CQID010	0.2115	0.4107	0.6250	0.4821	0.5179	0.1071	0.3036
CQID011	0.6869	0.7863	0.7417	0.8333	0.8256	0.7045	0.6821
CQID012	0.5737	0.6250	0.6696	0.6086	0.6280	0.5179	0.6161
CQID015	0.2500	0.4821	0.5000	0.7679	0.5714	0.3750	0.3393
CQID020	0.4641	0.5698	0.5002	0.5708	0.6188	0.5589	0.4054
CQID025	0.4038	0.7500	0.5357	0.8929	0.6250	0.0714	0.4643
CQID034	0.3269	0.3571	0.4286	0.4643	0.4643	0.1964	0.4286
CQID035	0.0577	0.5714	0.6786	0.8750	0.2500	0.3393	0.6607
CQID036	0.3077	0.3750	0.3214	0.5536	0.4107	0.3036	0.4107
Average	0.5556	0.5475	0.5556	0.6721	0.5457	0.3527	0.4790

Table 5

Fine-tuned model performance comparison across language models on validation dataset. Qwen2.5-VL-3B achieves the highest average performance.

Question	Qwen2-VL-2B	Qwen2-VL-7b	Qwen2.5-VL-3B	Qwen2.5-VL-7B	Gemma3-4B	Gemma3-12B	LLaMA-3.2-11B-Vision
CQID010	0.3600	0.2321	0.6607	0.2500	0.5714	0.2500	0.3571
CQID011	0.6747	0.3667	0.7634	0.3839	0.8360	0.8435	0.4976
CQID012	0.5567	0.5119	0.6726	0.5119	0.6577	0.5417	0.5506
CQID015	0.3200	0.0893	0.5357	0.2857	0.4107	0.8036	0.2857
CQID020	0.4560	0.2942	0.5126	0.3031	0.6253	0.6071	0.2990
CQID025	0.4600	0.6250	0.4821	0.3929	0.2500	0.0893	0.4107
CQID034	0.2600	0.1250	0.3571	0.0357	0.3036	0.3214	0.3750
CQID035	0.3200	0.1250	0.8214	0.7679	0.4821	0.5179	0.5714
CQID036	0.2200	0.2321	0.4107	0.5000	0.0714	0.1786	0.2500
Average	0.4030	0.2890	0.5796	0.3812	0.4676	0.4614	0.3997

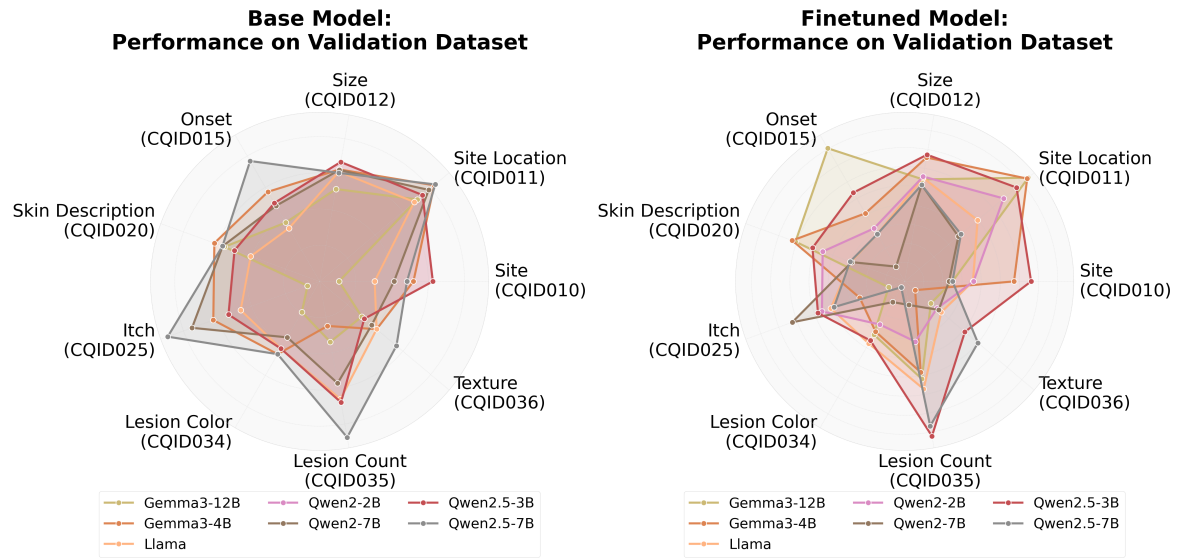


Figure 4: Performance Comparison of All Base and Fine-tuned Models. Radar plots depict model accuracy across individual diagnostic questions, highlighting subtle improvements for some models due to fine-tuning, while worsening the performance of other models. For example, the performance of the Gemma3-12B model improves greatly due to fine-tuning, while the performance of the Qwen2.5-7B noticeably degrades on most of the questions.

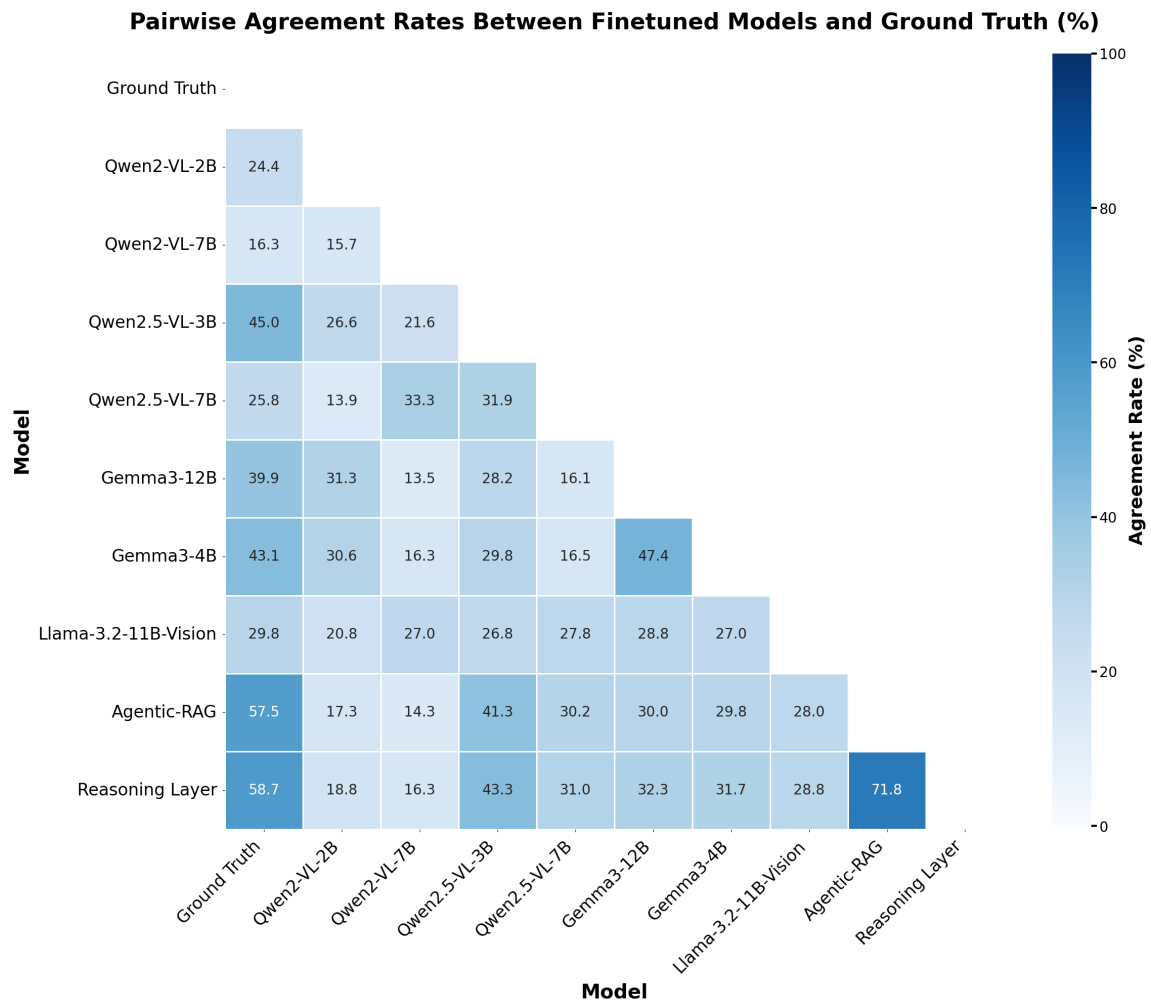


Figure 5: Pairwise agreement rates (%) among finetuned model predictions and ground truth labels. Higher values indicate greater consistency in labels.

Table 6: Reasoning Layer Output for Skin Description Task (CQID020) and Encounter 914.

Field	Content
Query Details	Query: What label best describes the affected area? Query Type: Skin Description Category: Skin Specific Clinical Context: Would brothers and sister here please help. These are my friend's hands. Has been to a few hospitals, but unable to get treated. Hope that people here can take a look to see what it is and what treatment is needed. Alternative medical treatment is also welcome. Thanks a lot. Options: raised or bumpy, flat, skin loss or sunken, thick or raised, thin or close to the surface, warty, crust, scab, weeping, Not mentioned
Finetuned Model Predictions	Qwen2-VL-2B: skin loss or sunken, warty, flat, Qwen2-VL-7B: raised or bumpy, thin or close to the surface, Qwen2.5-VL-3B: skin loss or sunken, warty, raised or bumpy, thick or raised, Qwen2.5-VL-7B: raised or bumpy, thick or raised, warty, Gemma-3-4B: crust, crust flat not mentioned, scab, , thin or close to the surface, warty, Gemma-3-12B: thick or raised, crust, raised or bumpy, LLaMA-3.2-11B: thick or raised, crust, raised or bumpy, scab, bumpy, warty
Reasoning	"The image analysis provides a detailed description of the affected areas across different sites (fingers, foot, ankle/lower leg). The 'SKIN_DESCRIPTION' section states that the morphology varies by location. - On the fingers, lesions are described as 'flat or slightly raised papules/confluent areas with dry skin and scaling'. This supports the options 'flat' and 'raised or bumpy'. - On the foot, lesions are described as 'discrete, raised papules... some crusted and excoriated'. This supports 'raised or bumpy', 'crust', and 'scab'. - On the ankle/lower leg, there are 'erythematous plaques with thick scaling, crusting, excoriations'. Plaques are elevated and often thickened, supporting 'raised or bumpy' and 'thick or raised'. 'Thick scaling' also supports 'thick or raised'. 'Crusting' supports 'crust' and 'scab'. Therefore, based on the evidence, the affected areas exhibit morphologies described by multiple labels from the options: flat, raised or bumpy, thick or raised, crust, and scab. These are all explicitly mentioned features of the affected skin across different sites."
Final Answer	raised or bumpy, flat, thick or raised, crust, scab

Table 7: Agentic RAG for Skin Description Task (CQID020) and Encounter 914.

Field	Content
Query Details	Query: What label best describes the affected area? Query Type: Skin Description Category: Skin Specific Clinical Context: Would brothers and sister here please help. These are my friend’s hands. Has been to a few hospitals, but unable to get treated. Hope that people here can take a look to see what it is and what treatment is needed. Alternative medical treatment is also welcome. Thanks a lot. Options: raised or bumpy, flat, skin loss or sunken, thick or raised, thin or close to the surface, warty, crust, scab, weeping, Not mentioned
Finetuned Model Predictions	Qwen2-VL-2B: skin loss or sunken, warty, flat, Qwen2-VL-7B: raised or bumpy, thin or close to the surface, Qwen2.5-VL-3B: skin loss or sunken, warty, raised or bumpy, thick or raised, Qwen2.5-VL-7B: raised or bumpy, thick or raised, warty, Gemma-3-4B: crust, crust flat not mentioned, scab, , thin or close to the surface, warty, Gemma-3-12B: thick or raised, crust, raised or bumpy, LLaMA-3.2-11B: thick or raised, crust, raised or bumpy, scab, bumpy, warty
Image and Clinical Context Analysis (Summarized)	Size: Lesions range widely by location—1–3 mm papules on the foot and fingers, up to several cm patches on the lower leg and fingers. Site Location: Involves dorsal/lateral fingers, medial malleolus and anterior distal lower leg/ankle, dorsal feet/toes, and interdigital web spaces. Some fingers are unaffected. Skin Description: Erythematous, dry, and scaly fingers; crusted plaques with punctate bleeding and excoriations on the lower leg/ankle; reddish-brown papules/nodules with central crusting on the foot/toes. Possible vesicles on some fingers. Lesion Color: Varies by site—mild erythema with whitish scales on fingers, yellowish-white scales/crusts and puncta on lower leg, reddish-brown papules and darker crusting on feet/toes. Lesion Count: Extensive involvement—diffuse finger changes, 5–10+ distinct lesions on lower leg/ankle, 20+ papules/nodules on foot/toes, clusters on fingers. Extent: Localized to fingers, lower legs, and feet as shown. Some hands partially unaffected. Full body extent not assessed. Texture: Consistently rough or bumpy—due to scaling and crusts (fingers, lower leg) and papules/nodules with crusting or excoriation (feet/toes). Onset Duration: Appears subacute to chronic, with signs of flaring. History of prior visits suggests long-standing or recurrent nature. Symptoms: Visual signs of pruritus include excoriations and crusts. Seeking treatment implies symptomatic discomfort. History: Previous unsuccessful treatment attempts across multiple hospitals.
Diagnosis Extractor	Diagnosis: Lichen Confidence: 0.70 Source: Image analysis
Retrieved Knowledge (Summarized)	Topic 1: Lichen planus overview Relevance Score: 2.47

Field	Content
	Key Information: – Affects skin, mouth, genitals, scalp, and nails; may appear in one or several locations – Can alter the appearance of fingernails and toenails – Often develops on the wrists but may involve other areas of the body – Not contagious and not a form of cancer Topic 2: Lichen planus diagnosis and treatment Relevance Score: 2.25 Key Information: – Diagnosis may include physical exam, skin biopsy, and blood tests – Treatment focuses on reducing symptoms and accelerating skin healing – Topical and oral corticosteroids, antihistamines, PUVA therapy, and retinoids are options – Mouth involvement may require different management approaches – Skin cases usually resolve within two years; recurrence is possible – Residual pigmentation changes may persist and require cosmetic treatments Topic 3: Neurodermatitis overview Relevance Score: 1.90 Key Information: – Characterized by intense, localized itching often affecting arms, legs, scalp, or groin – Itching may worsen during rest or nighttime and continue even without stress triggers – Persistent scratching can lead to thickened skin, bleeding, and secondary infections – Common signs include chronic itch, skin damage from scratching, and disturbed sleep – Diagnosis is based on clinical signs and ruling out other causes of chronic itch Topic 4: Lichen planus signs and symptoms Relevance Score: 1.61 Key Information: – Appears as shiny, firm, reddish-purple bumps often with white lines (Wickham’s striae) – May present as thick plaques, open sores, or scattered bumps on areas like wrists, back, or ankles – Can affect nails (causing splitting or thinning) and the scalp (causing irritation or scarring hair loss) – Oral form includes painful sores, redness, peeling gums, and white streaks or dots Topic 5: Neurodermatitis self-care strategies Relevance Score: 1.02 Key Information: – Daily management includes cool compresses, colloidal oatmeal baths, and antihistamines – Fragrance-free moisturizers and barrier protection reduce flare-ups – Covering affected areas and keeping nails short can minimize scratching damage – Heat, stress, and synthetic fabrics are common triggers to avoid

Field	Content
	– Consistent follow-up with dermatology improves long-term outcomes – Medical treatment is typically required for sustained symptom relief
Medical Knowledge Insights	"The retrieved medical knowledge discusses Lichen Planus and Neurodermatitis (Lichen Simplex Chronicus). Lichen Planus (Result 1, 4) can affect various sites including skin, nails, mouth, and has characteristic shiny, reddish-purple papules/plaques with white lines, often itchy. Neurodermatitis/LSC (Result 3, 5) is characterized by intense itching in one or two areas, leading to thickened, leathery skin from scratching, often on arms, legs, neck, scalp, groin. The visual findings of scaly/crusted patches/plaques on the lower leg (Image 2) and particularly the intensely pruritic-appearing, excoriated papules/nodules on the foot/toes (Image 3) are highly suggestive of Lichen Simplex Chronicus or a form of eczematous dermatitis driven by scratching. The diffuse scaling/dryness on other fingers (Image 1) could represent a different process or milder involvement. Image 5's potential vesicles could suggest dyshidrotic eczema. While classical Lichen Planus morphology (shiny, purple papules, Wickham's striae) is not prominent in these images, Lichen Planus can have varied presentations (hypertrophic LP, bullous LP) or co-exist with other conditions. The history of multiple failed treatments aligns with conditions like chronic eczema or LSC, which can be challenging to manage and often recur (Result 2, 3, 5). The intense itching visually evident is a key symptom of both LP and Neurodermatitis, prompting consideration of these diagnoses. Self-care measures for Neurodermatitis (Result 5) highlight the importance of breaking the itch-scratch cycle, which appears crucial based on the visual findings of excoriation."
Initial Reasoning	"The question asks for the best label(s) to describe the affected area based on the provided evidence. I need to review the 'SKIN_DESCRIPTION', 'TEXTURE', and 'SIZE' fields in the integrated evidence, as these directly describe the appearance of the lesions. The 'SKIN_DESCRIPTION' mentions:- Diffuse erythema, dryness, and scaling on fingers. Erythematous patches/plaques with yellowish-white scales, crusting, erosions, punctate bleeding, smaller papules, and excoriations on the lower leg/ankle.- Multiple discrete erythematous to reddish-brown papules/nodules, some with central crusting/excoriation on foot/toes. Small papules/vesicles on some fingers. The 'TEXTURE' is described as Roughness... due to scaling/dryness..., scales/crusts/erosions..., and papules/nodules with crusted/excoriated surfaces... Some fingers... may have bumpy texture from papules/vesicles. The 'SIZE' mentions papules/patches and papules/nodules; indicating elevated lesions. Evaluating the options: 'raised or bumpy': Supported by the mention of papules, nodules, vesicles, plaques, and bumpy texture. 'flat': Patches are mentioned, which are flat, but the description also includes scaling, crusting, papules, nodules, and plaques, which are not flat. This isn't the best overall description of the *affected area* which contains multiple lesion types. 'skin loss or sunken': Supported by the mention of erosions and excoriations, which represent skin loss. However, this describes a specific feature within some lesions, not the overall description of the raised/thickened areas. 'thick or raised': Supported by the mention of plaques and the medical knowledge linking findings to conditions like Neurodermatitis which cause skin thickening. Papules and nodules are also raised. 'thin or close to the surface': Not supported by the evidence. Thickening is implied by plaques and the likely diagnosis. 'warty': Not mentioned or suggested by the descriptions. 'crust': Explicitly mentioned as present on the lower leg/ankle and foot/toes. 'scab': Implied by the presence of crusting, erosions, and punctate bleeding. A scab is essentially a crust over a wound. 'weeping': Not explicitly mentioned as a prominent feature. 'Not mentioned': Several features are mentioned.

Field	Content
	Based on the evidence, the affected areas exhibit a variety of morphologies. Key features include raised lesions (papules, nodules, plaques), surface changes (scaling, crusting), and consequences of scratching (erosions, excoriations, bleeding, likely scabs). Therefore, multiple labels are appropriate to best describe the diverse appearance. 'raised or bumpy' is a good general term for the elevated lesions. 'thick or raised' captures the larger, possibly thickened plaques and the potential for chronic thickening. 'crust' is a direct description of a common surface finding. 'scab' is a very likely finding given the excoriations and bleeding leading to crust formation. All four of these options are well-supported by the integrated evidence describing different aspects of the affected areas."
Initial Results	Predicted Answer: raised or bumpy, thick or raised, crust, scab (Confidence: 0.74). Uncertainty factors: – "Potential overlap or redundancy between 'raised or bumpy' and 'thick or raised'." – "Inferring the presence of 'scabs' from 'crusting, erosions, and punctate bleeding' rather than explicit mention." Counterfactual: "If the evidence only described 'flat erythematous patches with scaling', then 'flat' would be a better descriptor and 'raised or bumpy', 'thick or raised', 'crust', and 'scab' would be less appropriate or incorrect. If the evidence explicitly stated the absence of crusting or excoriations, then 'crust' and 'scab' would not be selected." Knowledge Contribution: "The medical knowledge about Neurodermatitis (Lichen Simplex Chronicus) reinforces the relevance of 'thick or raised' by explaining that intense scratching can lead to thickened skin (lichenification), which aligns with the description of plaques and the visually evident excoriations. This adds weight to 'thick or raised' as a key descriptor beyond just elevated lesions."
Reflection + Revision	Reflection: "The initial reasoning correctly identified the key evidence fields ('SKIN_DESCRIPTION', 'TEXTURE', 'SIZE') and extracted relevant descriptions of the lesions (scaling, crusting, erosions, papules, nodules, plaques, bumpy texture). It appropriately linked these findings to several of the options, correctly identifying support for 'raised or bumpy', 'thick or raised', and 'crust'. The inference of 'scab' from crusting, erosions, and bleeding is also a reasonable interpretation, as scabs are composed of dried blood and serum, often forming over erosions or excoriations. The reasoning correctly dismissed 'flat' as not representative of the entire affected area, and correctly dismissed 'thin or close to the surface', 'warty', and 'weeping' due to lack of evidence. However, the reasoning for excluding 'skin loss or sunken' was weak. While it correctly noted that erosions and excoriations (representing skin loss) are features *within* some lesions, it failed to fully appreciate that these features are widespread and significant descriptors of the affected area, particularly on the lower leg and foot/toes. The question asks to describe the 'affected area', which includes all observed pathologies, not just the raised or thickened components. Skin loss due to scratching is a prominent characteristic visually evident in the provided images and described in the evidence, making 'skin loss or sunken' a valid and important descriptor of parts of the affected area." Reasoning Gap: "The primary reasoning gap was the assumption that a descriptor must apply uniformly to the entire 'affected area' to be selected. The affected area encompasses diverse morphologies across different sites. Descriptors like 'skin loss or sunken', 'crust', 'scab', 'raised or bumpy', and 'thick or raised' all describe different, co-existing features within the overall affected region. The initial reasoning correctly applied this principle to select multiple 'raised' descriptors and 'crust', but inconsistently applied it by dismissing 'skin loss' because it didn't describe the *raised* areas." Revised Answer: raised or bumpy, skin loss or sunken, thick or raised, crust, scab.

Field	Content
	Revised Confidence: 0.90 Revised Reasoning: "The question asks for the best label(s) to describe the affected area based on the integrated evidence. I need to perform a detailed analysis of the visual descriptions ('SKIN_DESCRIPTION', 'TEXTURE', 'SIZE') across the different affected sites mentioned (fingers, lower leg/ankle, foot/toes) and consider how these features relate to the provided options. The reflection highlighted the importance of considering the diversity of lesions within the 'affected area' and ensuring that features like skin loss from excoriation are not overlooked." The 'SKIN_DESCRIPTION' reveals a heterogeneous presentation across body sites: - Fingers: diffuse erythema, dryness, scaling, possibly small papules/vesicles. - Lower leg/ankle: erythematous patches/plaques with scales, crusting, erosions, punctate bleeding, smaller papules, excoriations. - Foot/toes: discrete papules/nodules, some with central crusting/excoriation. The 'TEXTURE' is described as rough due to scaling, crusts, erosions, papules, and nodules, and potentially bumpy from papules/vesicles. The 'SIZE' includes papules (1-3mm), nodules, patches, and plaques (up to several cm). Evaluating the options based on these descriptions: **raised or bumpy**: This is strongly supported. Papules, nodules, and plaques are explicitly mentioned, all of which are elevated lesions. The texture is described as rough and potentially bumpy from papules/vesicles. This applies to lesions on fingers (papules/vesicles), lower leg/ankle (papules, plaques), and foot/toes (papules/nodules). **flat**: Patches are mentioned on the lower leg/ankle. Patches are typically flat areas of altered color or texture. However, these patches have additional features like scales, crusting, erosions, and papules on them, meaning the surface is not uniformly flat. The diffuse erythema on fingers might be primarily flat, but is associated with scaling. Given the widespread presence of raised lesions (papules, nodules, plaques) and surface irregularities (scaling, crusting, erosions, excoriations), 'flat' is not the best overall descriptor for the 'affected area' encompassing all these sites and morphologies. **skin loss or sunken**: This is strongly supported. 'Erosions' and 'excoriations' are explicitly mentioned on the lower leg/ankle and foot/toes. Erosions are superficial skin loss. Excoriations are deeper abrasions caused by scratching, also representing skin loss. The term 'skin loss' is directly applicable to these features. While the lesions aren't described as 'sunken' (which often implies atrophy or subcutaneous volume loss), the 'skin loss' part of the option is clearly present and significant, particularly in areas where scratching is severe (visually evident and supported by 'SYMPTOMS'). **thick or raised**: This is strongly supported. Papules, nodules, and plaques are all 'raised'. Plaques on the lower leg/ankle suggest thickened skin. The medical knowledge about Neurodermatitis/Lichen Simplex Chronicus ('MEDICAL_KNOWLEDGE_INSIGHTS') reinforces that chronic scratching leads to skin thickening (lichenification), which aligns with the description of plaques and the visual evidence of intense pruritus (excoriations). Thus, 'thick' is a highly relevant descriptor for some areas, particularly the lower legs, and 'raised' applies across all sites with papules/nodules/plaques. **thin or close to the surface**: 'Thin' is not supported by any evidence; in fact, thickening is suggested. Some lesions like erosions or superficial papules are 'close to the surface', but 'thin' as a descriptor for the skin texture is inaccurate. **warty**: There is no description of warty appearance (i.e., papillomatous, verrucous surface). **crust**: This is explicitly mentioned as a feature on the lower leg/ankle and foot/toes. This is a direct and accurate descriptor of a surface characteristic in the affected area. **scab**: This is strongly implied. The presence of 'crusting', 'erosions', and 'punctate bleeding' indicates areas where the skin surface has been broken (erosions, excoriations) and dried blood, serum, and exudate have formed a crust.

Field	Content
	A scab is precisely such a formation over a wound or abrasion. This is a very likely finding given the described features. 9. **weeping**: Not explicitly mentioned as a prominent feature. While some serous exudate leading to crusting might occur, 'weeping' typically implies active discharge, which is not described. Multiple options are necessary to accurately describe the varied affected areas. The analysis supports 'raised or bumpy', 'skin loss or sunken', 'thick or raised', 'crust', and 'scab' as valid descriptors of different aspects of the observed lesions and affected regions." Knowledge utilization assessment: "Medical knowledge about conditions like Lichen Simplex Chronicus (Neurodermatitis) was integrated to understand the significance of visually evident features like intense excoriation and described findings like plaques. This knowledge reinforces that chronic scratching leads to skin thickening (lichenification), validating 'thick or raised' as a descriptor, and also explains why skin damage (erosions, excoriations, leading to skin loss, crusts, and scabs) is a prominent feature in such pruritic conditions. This contextual understanding strengthens the evidence supporting descriptors like 'thick or raised', 'skin loss or sunken', 'crust', and 'scab' in the context of the patient's history and visual presentation."
Final Answer	"raised or bumpy, skin loss or sunken, thick or raised, crust, scab" Confidence: 0.90