A synthetic data generation framework for scalable and resource-efficient medical AI assistants

Abdurrahim Yilmaz[1], Furkan Yuceyalcin[2], Rahmetullah Varol[3], Ece Gokyayla[4], Ozan Erdem[5], Donghee Choi[1], Ali Anil Demircali[1], Gulsum Gencoglan[6,*], Joram M. Posma[1,*], Burak Temelkuran[1,*]

[1] Division of Systems Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College London
[2] Mechatronics Engineering Department, Yildiz Technical University
[3] Department of Business Administration, Universitat der Bundeswehr München
[4] Department of Dermatology and Venereology, Usak Research and Training Hospital
[5] Department of Dermatology and Venereology, Istanbul Medeniyet University
[6] Department of Dermatology and Venereology, Medicana Atakoy Hospital
[*] These authors jointly supervised this work: Burak Temelkuran (b.temelkuran@imperial.ac.uk), Joram M. Posma (j.posma11@imperial.ac.uk), Gulsum Gencoglan (ggencoglan@gmail.com)

Abstract

Large language models (LLMs) show promise in the medical field but remain limited by high computational costs and privacy needs. We present SCALEMED (Scalable Clinical Assistants and LEarning for MEDicine), a data-centric framework designed to train and use specialized medical models on standard hardware. AnnotatorMed, an open-source annotation tool supporting local data labelling, is also developed thereby reinforcing privacy. For dermatology — a visually demanding field — SCALEMED integrates open-access (OA) data, PubMed (PM) literature, and instruction-following (IF) to create DermaSynth, a comprehensive dataset of dermatological images and reports via knowledge transfer with 1.2 million samples. We demonstrate DermatoLlama, a vision-language model that inherits high-level reasoning using DermaSynth yet remains resource-efficient for local deployment. By unifying open-source tools, privacy-focused design, and modular training approaches, SCALEMED provides adaptable AI solutions in resource-limited clinical settings, such as our exemplar of dermatology. This work highlights a practical path for healthcare institutions to adopt LLM-driven decision support systems without the burdens of large-scale architectures or external data exposure by reaching the success of state-of-the-art (SOTA) vision LLMs.

Introduction

Large language models (LLMs) are increasingly recognized as transformative tools in healthcare, enabling the creation of explainable analysis, structured diagnoses, and personalized treatment plans [1–4]. They have also demonstrated promise in providing multi-modal solutions such as interpreting radiological images, augmenting clinical decision-making, and streamlining workflows [5–7]. However, despite these advancements, LLM applications in medicine still face significant barriers in widespread adoption. This is due to the limited availability of high-quality and diverse datasets [8], stringent privacy regulations [9], and the high computational and staff costs [10, 11] associated with training and deploying LLMs in clinical settings. Specifically, these constraints cause significant challenges for the adaptation of LLMs to real-world and resource-constrained environments. In this context, resource-efficient and privacy-preserving solutions allowing scalability to multiple fields are essential for the effective integration of LLMs into healthcare practice.

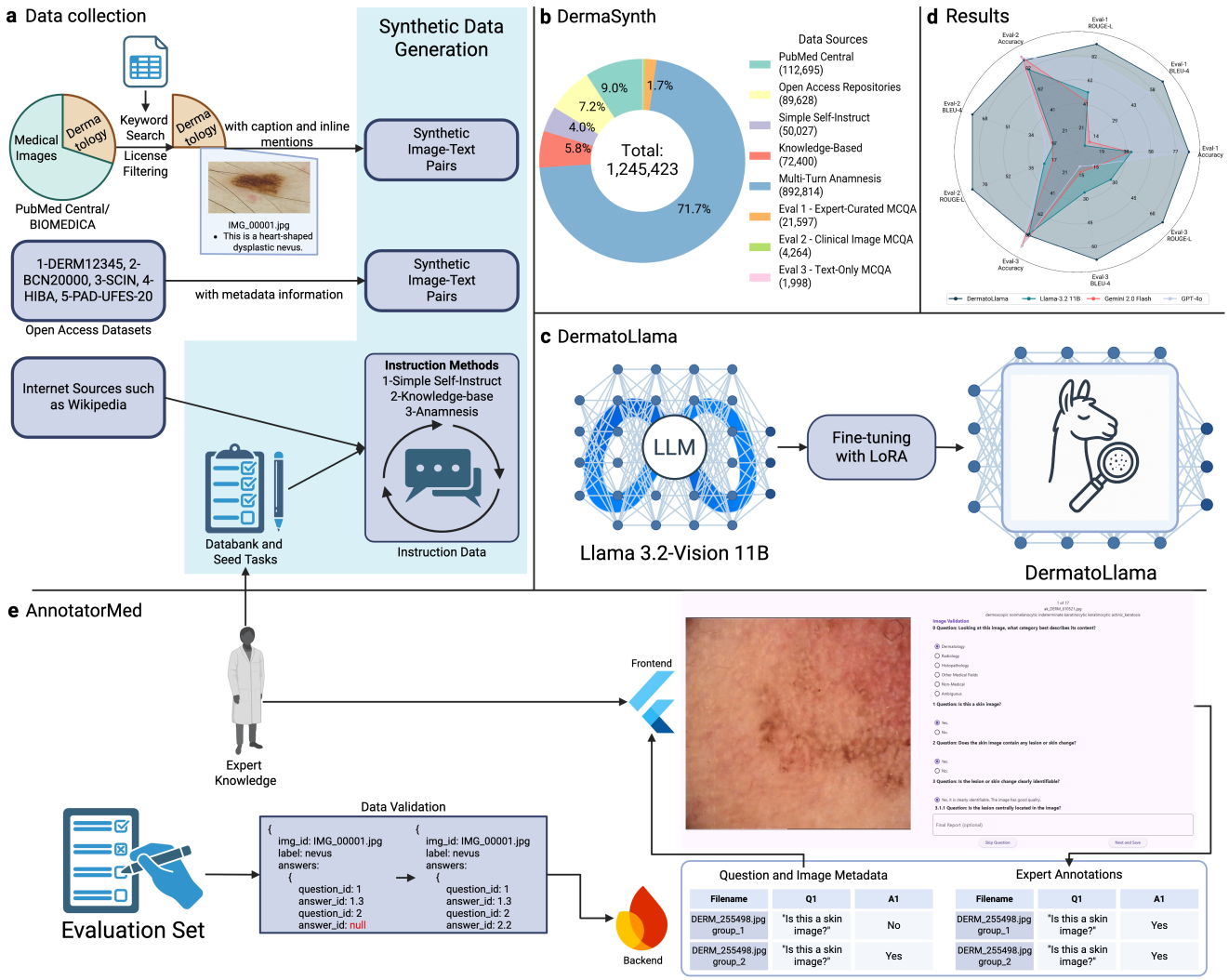

Figure 1: **Overview of SCALEMED framework.** **a,** Data collection pipeline of DermaSynth—1.2 million data samples based on PubMed, open access datasets and instruction-follow methods. **b,** DermaSynth data size and its data sources. **c,** DermatoLlama was fine-tuned on Llama-3.2-11B-Vision using LoRA optimization method. **d,** DermatoLlama can reach GPT-4o, Gemini 2.0 Flash in accuracy, and can surpass all model in report generation for all evaluation sets. **e,** AnnotatorMed is a free and open source tool for data annotation in clinical context.

Building task-focused artificial intelligence (AI) assistant systems (an 'agent' or 'multi-agents') and generating synthetic data with or without ground-truth have emerged as powerful strategies for overcoming the high computational and extensive data demands of LLMs in healthcare [12–14]. By transferring knowledge from a large teacher model to a lighter and efficient student model, knowledge transfer reduces the resources needed for training and inference—an approach particularly useful in settings with limited computing, budgetary constraints, and labour force. Collaborative AI assistants ('multi-agents'), in turn, distribute complex tasks among specialized assistants that communicate and cooperate, allowing medical AI to handle diverse inputs (such as images and texts) while maintaining flexibility and scalability [15–18]. Despite these promising methodologies, current evaluation practices and data-handling approaches rarely harness LLMs to their

full potential, focusing instead on single-assistant scenarios or requiring large-scale hardware access [19]. Moreover, privacy concerns and data scarcity remain pressing issues. Many organizations are not allowed to use AI models hosted by external servers which requires sharing confidential medical data. In addition, researchers need access to high quality processed data to build sufficiently representative datasets to capture the full spectrum of patient variation. Consequently, the field lacks a unified framework that incorporates knowledge transfer, multi-assistant designs, and multi-modal synthetic data generation to address these real-world constraints. Such an integrated solution could accelerate the deployment of LLM-driven healthcare tools, enabling robust performance across medical specialties—even in resource-constrained settings—while safeguarding patient confidentiality and optimizing resources needed for institutional usage.

To address these challenges in developing robust, privacy-preserving, and scalable AI systems for healthcare, we propose a data-centric framework, called SCALEMED (Scalable Clinical Assistants and LEarning for MEDicine). In contrast to traditional approaches that rely on tightly controlled datasets and static evaluation formats, SCALEMED integrates synthetic data generation, leveraging open-access (OA) databases, and PubMed (PM) literature to create a diverse, representative dataset of medical information, with a specific use case in dermatology. This dataset is enriched through flexible instruction methods [20–22], ranging from AI-generated clinical questions using high quality refernces (self-instruct [20]) to medical knowledge-based prompts and anamnesis-style narratives. These flexible instruction methods allow additional instruction methods to be integrated, exposing vision LLMs to a wide array of clinical scenarios. SCALEMED also highlights critical license considerations, guiding researchers on the safe and compliant utilization of publicly available data. Within SCALEMED, users can generate synthetic evaluation data optionally using state-of-the-art (SOTA) LLMs for various diagnostic scenarios, aided by customizable code scripts and open-source annotation tool, AnnotatorMed. To enhance computational efficiency, SCALEMED can be integrated techniques like Low-Rank Adaptation (LoRA) [10] and Quantized LoRA (QLoRA) [11] during model fine-tuning. These techniques allow for the efficient adaptation of OA models like Llama and its vision variants [23], significantly reducing the computational load without sacrificing model performance [10]. It further refines these models through knowledge transfer, while respecting patient privacy through local data annotation and model development on ordinary computing hardware. An iterative, modular design ensures the framework is adaptable, explainable, and scalable, enabling multi-agentic systems where specialized agents can collaborate. By consolidating these elements—synthetic data generation, model fine-tuning, open-source tools, expert annotation, privacy safeguards and flexible deployment - SCALEMED aims to deliver a transparent, ethical, and high-performing ecosystem for real-world medical AI applications. By applying these methods, SCALEMED ensures that smaller, domain-specific models can be fine-tuned and deployed on cost-efficient private servers, supporting scalable and specialized AI agents in clinical scenarios.

Here we demonstrate the SCALEMED framework for dermatology, a domain that lacks sufficient diverse and high-quality datasets necessary for advanced reasoning [24,25], which this framework aims to solve. Applying our framework to dermatology, we created DermaSynth, a synthetic dataset of approximately 1.2M samples assembled by scraping 56,258 images from PubMed articles. We enriched these synthetic data created using clinical instructions (instruction-followings (IF)) through 193 seed tasks for simple self-instruction, 40 seed tasks for medical knowledge-based prompts, and 137 seed tasks reflecting multi-turn, anamnesis-style doctor–patient interactions—collectively serving as knowledge transfer from domain sources to the model. These methods help replicate real-world dermatological challenges at scale, allowing the vision–language model, DermatoLlama, to acquire both breadth and depth of clinical insight. A code-driven eval-

uation pipeline generates synthetic question–answer (QA) pairs, including image-based queries, enabling systematic performance assessment without reliance on sensitive patient data. AnnotatorMed, an open-source annotation tool developed within this framework, preserves privacy by supporting local development on standard hardware, while its modular, iterative, and hierarchical design facilitates smaller, cost-efficient deployments. Early results demonstrate DermatoLlama's capacity to handle a wide array of dermatological scenarios with transparency and scalability.

Results

SCALEMED Framework Implementation and DermaSynth Generation

We successfully implemented the SCALEMED framework (Fig. 1) to generate a large-scale, multifaceted synthetic dataset for dermatology, named DermaSynth. For DermaSynth, we generated approximately 1.2M synthetic samples. This dataset integrates visual question-answer (VQA) pairs derived from metadata-rich OA datasets (n = 26,888 QA pairs of 13,444 clinical images and n = 62,740 QA pairs of 31,370 dermoscopic images) (such as DERM12345 [26]) and the clinically relevant BIOMEDICA dataset (n = 37,565 sample and n = 112,695 image-text pairs) [27] that is based on OA PubMed articles. For the synthetic instruction-following dataset, we firstly designed seed tasks to cover all possible clinical scenarios. Diverse text-based data were generated via self-instruction as: 1) IF QA (n = 193 seed tasks, n = 50,027 synthetic samples), 2) knowledge-base QA (n = 40 seed tasks, n = 72,400 synthetic samples) and 3) anamnesis case studies (n = 137 seed tasks, n = 892,814 samples of n = 9,963 synthetic cases). We also designed three evaluation sets to evaluate VLMs for dermoscopic image analysis within a multi-turn conversation (Eval1, n = 21,597), clinic image analysis with report generation based on PubMed images (Eval2, n = 4,264), and common dermatology knowledge (Eval3, n = 1,998). This data-centric approach provided the baseline datasets for training and evaluating a specialized dermatology VLM.

DermatoLlama Performance on Benchmark Evaluations

We evaluated the performance of our fine-tuned models, DermatoLlama (Llama-3.2-11B-Vision fine-tuned on DermaSynth), against baseline open-source VLMs (base Llama-3.2-11B-Vision) and state-of-the-art models (GPT-4o - knowledge cutoff: October 2023, Gemini 2.0 Flash - knowledge cutoff: August 2024 [28]). The comparison was conducted across three distinct benchmarks (Eval1, Eval2, Eval3) targeting different aspects of dermatological reasoning and knowledge.

Furthermore, we conducted targeted evaluations to assess the quality and effectiveness of each distinct synthetic data generation pipeline comprising DermaSynth. We trained separate model instances by progressively adding more data (10k (OA), 50k (OA), 100k (OA), 200k (OA+PM), all samples (OA+PM+IF)) and evaluated them on their respective evaluation benchmark (Eval1, Eval2, Eval3). This targeted analysis validates the contribution of each data generation strategy. Detailed results for both the final DermatoLlama model trained on different sample sizes and SOTA models are presented in Table 1 and Fig. 2.

Evaluation on Eval1: Multi-Turn Dermoscopic Image Analysis

We designed a hierarchical question set for two-step pattern analysis [29] to guide systematic diagnosis. This question set contains 93 root questions and 301 total questions spanning five clinical-dermoscopic subsections (basic clinical context, clinical context, dermoscopic context, melanocytic

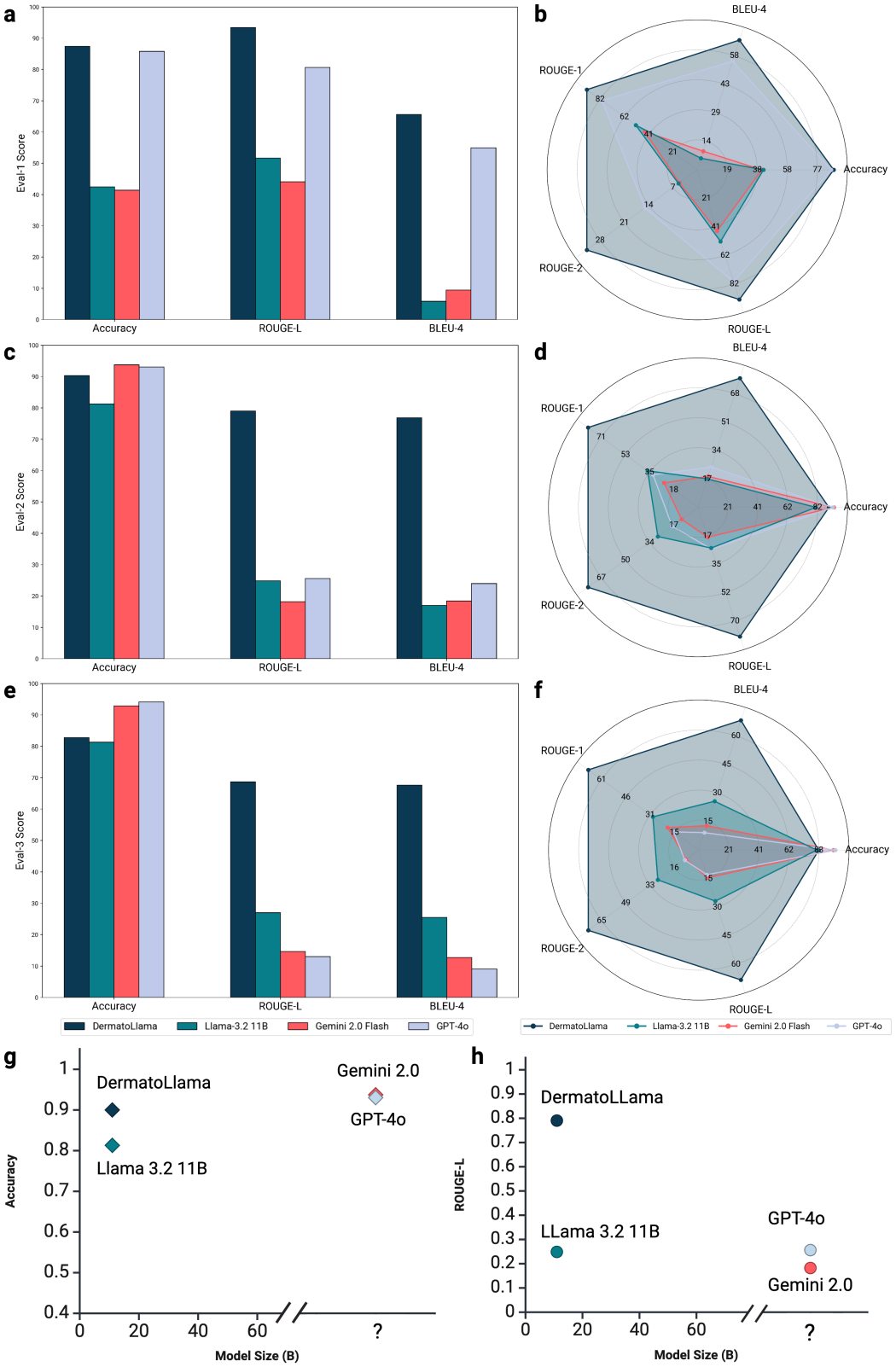

Figure 2: **Comparison of results. a,b,** Eval1 - classification and text performance of DermatoLlama. **c,d,** Eval2 - classification and text performance of DermatoLlama. **e,f,** Eval3 - classification and text performance of DermatoLlama. **g,h,** Eval 2 - Scatter plots of accuracy and ROUGE-L versus parameter count showcase DermatoLlama's -small-but-mighty- profile—delivering large-model performance from a compact footprint.

analysis, non-melanocytic analysis). Eval1 comprises 225 expert-annotated images from the test split of DERM12345 with single-choice, multi-choice and open-ended items that require chained clinical reasoning. The DermatoLlama model (OA+PM+IF) demonstrated significantly improved performance compared to the baseline Llama model. It reached an accuracy (Acc) of 0.87 and text-generation scores of BLEU-4 = 0.66, ROUGE-1 = 0.94, ROUGE-2 = 0.32 and ROUGE-L = 0.93. Relative to the baseline Llama model (Acc = 0.42, BLEU-4 = 0.06, ROUGE-1 = 0.52, ROUGE-2 = 0.05, ROUGE-L = 0.52). DermatoLlama also surpassed GPT-4o (Acc = 0.86, BLEU-4 = 0.55, ROUGE-1 = 0.81, ROUGE-2 = 0.15, ROUGE-L = 0.81) with better accuracy and report generation, while eclipsing Gemini 2.0 Flash (Acc = 0.41, BLEU-4 = 0.09, ROUGE-1 = 0.44, ROUGE-2 = 0.05, ROUGE-L = 0.44) across all metrics. Example reports of ground truth, DermatoLlama and base Llama model is presented in Fig. 3. The other variants of DermatoLlama (trained on different sample sets, OA, OA+PM, OA+PM+IF) is well above the baseline/SOTA models and confirming that converting clinical data into structured descriptions yields high-fidelity supervision. The additional data gathered from PubMed and instruction-following approaches demonstrates their complementary benefit of combining diverse data sources.

Evaluation on Clinical Multiple-Choice VQA (Eval2)

The Eval2 benchmark tested the models' ability to answer multiple-choice questions based on mostly clinical images and associated text from the BIOMEDICA dataset / PubMed Central [27]. DermatoLlama - Full achieved 0.91 accuracy with BLEU-4 = 0.77, ROUGE-L = 0.79. This model surpassed the original Llama model (Acc = 0.81, BLEU-4 = 0.17, ROUGE-L = 0.25). Although GPT-4o (0.93 accuracy) and Gemini 2.0 Flash (0.94 accuracy) have higher classification accuracy, however DermatoLlama outperforms both based on BLEU-4 and ROUGE-L, indicating superior report quality. The DermatoLlama - OA+PM (Acc = 0.89, BLEU-4 = 0.62, ROUGE-L = 0.65) and DermatoLlama - OA (Acc = 0.62, BLEU-4 = 0.61, ROUGE-L = 0.60) affirms the usefulness of PM-derived synthetic QA and showing that broader data exposure (full model) yields a further accuracy and BLEU-4. These results show that PM-derived synthetic data, which is based on high-trust source, significantly improves the model's ability to interpret dermatological figures in the clinical literature and that wider data exposure further enhances both diagnostic accuracy and language quality [30].

Evaluation on Text-Only Dermatology Knowledge MCQs (Eval3)

To assess the transfer of textual dermatological knowledge without images, we used the Eval3 benchmark, comprising text-only multiple-choice question answers (MCQA). DermatoLlama obtained 0.83 accuracy, BLEU-4 = 0.68 and ROUGE-L = 0.69, surpassing the baseline Llama model (Acc = 0.81, BLEU-4 = 0.26, ROUGE-L = 0.27) by better accuracy and BLEU-4. While GPT-4o (Acc = 0.94) and Gemini 2.0 Flash (Acc = 0.93) lead in classification, their report generation scores remain low (BLEU-4 $\leq$ 0.12; ROUGE-L $\leq$ 0.15), highlighting DermatoLlama's superior explanations for clinicians. This high success compared to the baseline confirms that the self-instruction strategies successfully generated a diverse and accurate dermatological knowledge base, effectively embedding domain-specific textual understanding into the language component of the VLM.

Resources Released

We make the following assets publicly available:

Data / Model	Acc	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
Eval1 - 225 samples from test split of DERM12345					
Raw	0.4244	0.0584	0.5201	0.0537	0.5163
10k - OA	0.7635	0.3430	0.7986	0.2633	0.7973
50k - OA	0.7874	0.3858	0.8197	0.2710	0.8186
100k - OA	0.8146	0.3746	0.7768	0.2385	0.7760
200k - OA + PM	0.8582	0.5304	0.9195	0.3098	0.9187
Full - OA + PM + IF	**0.8742**	**0.6559**	**0.9349**	**0.3160**	**0.9344**
Gemini 2.0 Flash	0.4140	0.0941	0.4441	0.0510	0.4405
GPT-4o	0.8580	0.5495	0.8066	0.1498	0.8065
Eval2 - PubMed Test Split					
Raw	0.8102	0.1699	0.3693	0.2785	0.2487
10k - OA	0.7085	0.2576	0.3365	0.1236	0.3259
50k - OA	0.5229	0.5890	0.6458	0.5870	0.6228
100k - OA	0.6203	0.6058	0.6572	0.5969	0.6348
200k - OA + PM	0.8874	0.6226	0.6797	0.6153	0.6542
Full - OA + PM + IF	0.9093	**0.7686**	**0.8056**	**0.7634**	**0.7904**
Gemini 2.0 Flash	**0.9374**	0.1842	0.2482	0.1142	0.1822
GPT-4o	0.9303	0.2399	0.3293	0.1796	0.2561
Eval3 - LLM-generated questions					
Raw	0.8133	0.2546	0.2893	0.2753	0.2699
10k - OA	0.7387	0.6122	0.6290	0.6951	0.6066
50k - OA	0.6597	0.6382	0.6642	0.7199	0.6481
100k - OA	0.6491	0.6420	0.6689	0.7198	0.6529
200k - OA + PM	0.7978	0.6616	0.6777	0.7267	0.6715
Full - OA + PM + IF	0.8273	**0.6764**	**0.6973**	**0.7432**	**0.6866**
Gemini 2.0 Flash	0.9284	0.1271	0.1962	0.0859	0.1463
GPT-4o	**0.9414**	0.0910	0.1584	0.0930	0.1304

Table 1: Classification and report generation performance of DermatoLlama with different variants and SOTA models, GPT-4o and Gemini 2.0 Flash. OA: Open access images with QA pairs, PM: Open access PubMed images with QA pairs, IF: Text-based instruction-following samples. The highest metric across all models per evaluation set is given in **bold**, and second highest is underlined.

- **DermaSynth** – a dataset of approximately **multimodal 1,2 million** synthetic data points around 250k image–text pairs (clinical + dermoscopic) in both Hugging Face Datasets and JSON formats.
- **DermatoLlama 1.0** – an 11B parameter, long-context (2,048 tokens) vision–language model fine-tuned with LoRA adapters on a DermaSynth mixed real + synthetic dermatology set to process longer clinical texts and images.
- **AnnotatorMed** – an open-source, locally deployable, and user-friendly web and mobile application. AnnotatorMed is a Flutter application with a Firebase backend for hierarchical data annotation, clinician test, and optional reinforcement learning from human feedback (RLHF) [31] logging, an approach to improve AI performance with relatively small feedbacks.
- **SCALEMED Python Framework** – end-to-end utilities (data scraping, synthetic generation, fine-tuning, evaluation) and a Colab notebook illustrating how to adapt the pipeline to any medical specialty.

All materials are hosted at **https://github.com/DermaVLM**; the model checkpoint is mirrored on Hugging Face at **https://huggingface.co/DermaVLM**. Use of the dataset, model, and code is restricted to **non-commercial research**, in accordance with the licensing terms documented in the repository.

Discussion

Deploying large language models (LLMs) in clinical practice remains challenging due to limited data availability, stringent privacy regulations, patient safety concerns, and the high computational cost of training/using these models. For many institutions, hosting extremely large-scale medical models [32, 33] and frameworks [34–36] are unfeasible because of constrained infrastructure and budgets. There is a critical need for resource-efficient approaches and integrated frameworks that provide succesful AI models within typical hospital environments.

In this study, we designed an integrated framework, SCALEMED, which is a modular, data-centric workflow for developing medical vision language models. SCALEMED combines synthetic data generation, open-source tools for data annotation, and large-scale fine-tuning in a single, scalable, end-to-end pipeline (Fig. 1). Here, we showed that (i) using synthetic data with targeted clinical instructions improved both classification and report-generation success; (ii) each synthetic data source—open-access images, PubMed figures, and self-instruction—contributes distinct and additive gains, and (iii) DermatoLlama's full configuration matched or surpassed SOTA models, GPT-4o and Gemini 2.0 Flash, on report generation metrics and achieved comparable classification accuracy. DermatoLlama surpasses SOTA models for the evaluation set, Eval1 which is published after the release of SOTA models used in this study. For images that may already be represented within their training dataset of these SOTA models (Eval2), DermatoLlama has superior text metrics and attains comparable accuracy. When evaluated on questions generated by Gemini 2.0 Flash (Eval3), the SOTA models retain a modest advantage, yet DermatoLlama still attains an accuracy of 83%. Across all three evaluation sets, it achieves the highest scores on every report-generation metric, consistently outperforming GPT-4o, Gemini 2.0 Flash, and other baselines 2.

Fine-tuned DermatoLlama with smaller sample sizes (10k–50k) had not yet converged and therefore generated detailed, multi-sentence outputs that inflated amount of generated text and

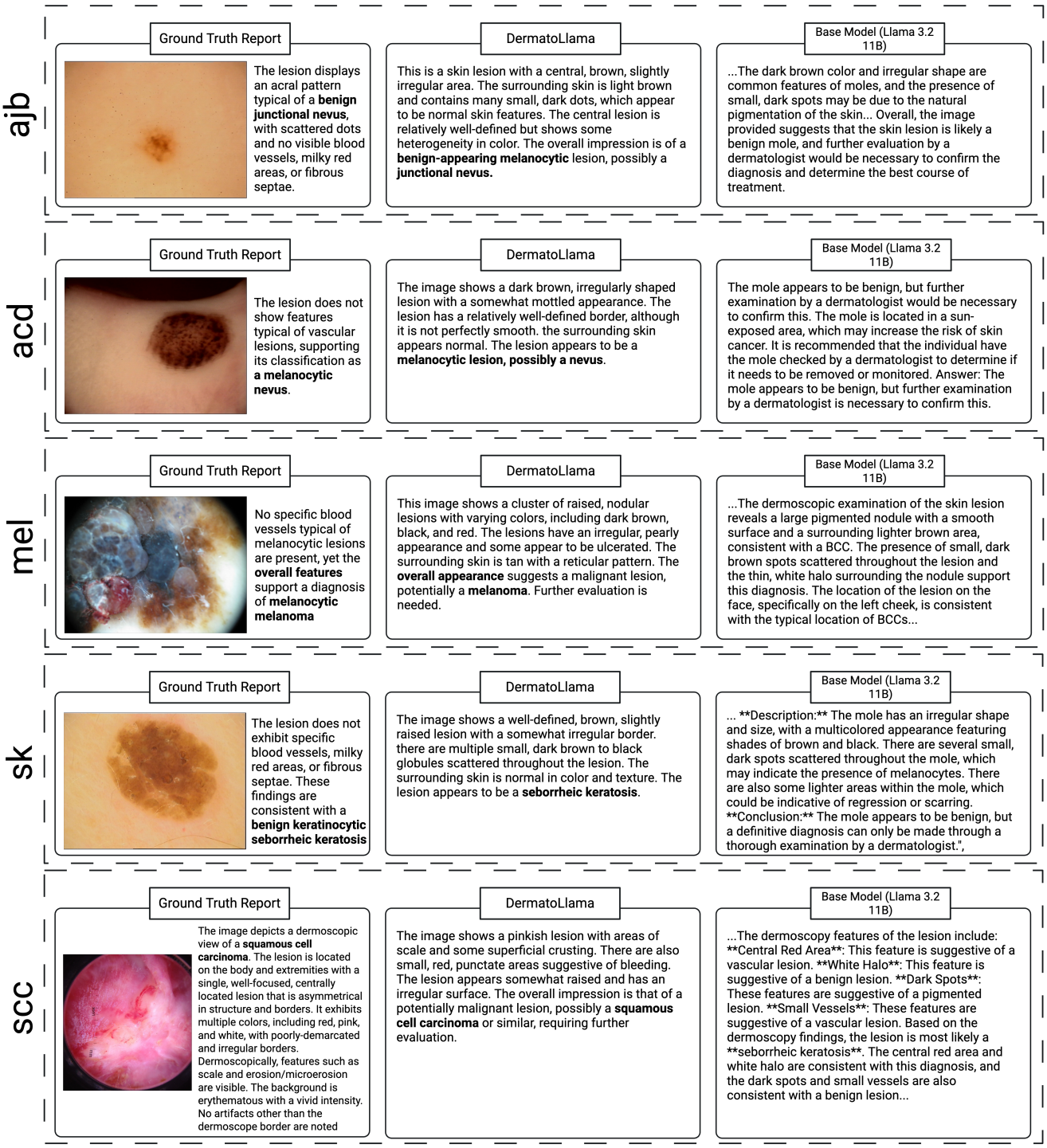

Figure 3: **Example reporting results.** Dermoscopic cases—ajb (acral junctional banal nevus), acd (acral compound dysplastic nevus), mel (melanoma), sk (seborrheic keratosis), scc (squamous cell carcinoma)—with the ground-truth report (left), DermatoLlama output (centre), and base Llama output (right). DermatoLlama descriptions align more closely with the reference than the base Llama model.

depressed length-sensitive metrics such as BLEU and ROUGE. DermatoLlama trained with all samples learned to output only the clinically relevant answer, eliminating unnecessary text, and producing improved report generation scores. BLEU-4 increased by 11.2× over the base model and 6.97× over Gemini on Eval 1; by 4.5× (base) and 4.17× (Gemini) on Eval 2; and by 2.65× (base) and 5.32× (Gemini) on Eval 3. By reducing unnecessary explanations, training and inference become more compute efficient, critical for hospitals with limited hardware resources. This focused output also simplifies human review during validation and annotation, enabling rapid feedback loops and more precise expert corrections.

In our evaluations, we observed a consistent improvement in performance as progressively larger and more diverse datasets were incorporated during fine-tuning. We increased the data size from thousands of open-access image–text pairs to hundreds of thousands of enriched image-text pairs with PubMed contexts and instruction-following samples. Starting from an untrained baseline, each increase in data volume yielded higher evaluation metrics. The 50k, 100k, and 200k configurations all showed marked gains over earlier training stages, while the Full dataset achieved the best overall scores. DermatoLlama performance plateaued around 200k samples; however, supplementing with synthetic data significantly broadened its effective knowledge base and boosted the success of DermatoLlama. These findings underscore the importance of multi-source data integration enhanced with synthetic data. Instruction-focused samples, introduced at the million scale, substantially improved the model's capacity to handle open-ended queries, refining both diagnostic accuracy and report quality.

A vital insight from our dermatology use case is the importance of high-quality synthetic datasets. In settings where real-world data is scarce or cannot be easily shared due to privacy concerns, synthetic data can fill gaps by simulating a broad range of clinical presentations. Our experiments reinforced that the diversity and representativeness of synthetic samples are important; however, synthetic data must be supported by clinically grounded data like PubMed Central. In addition, we found that the creation of seed tasks and databanks can be rapidly achieved with minimal but strategic clinician input, streamlining the generation of high-quality synthetic data. This reduces the overall burden on domain experts while ensuring clinically relevant outputs.

SCALEMED is designed to produce synthetic training data by using open-access data and user/expert-provided seed tasks or prompts to create clinically relevant image–text pairs and multi-turn conversation sets. SCALEMED allows human experts to refine this enriched dataset through a user-friendly annotation interface, AnnotatorMed, correcting or adjusting synthetic data and model outputs as needed. Subsequently, transfering knowledge from larger, established models ('teacher') into smaller, resource-efficient models ('student'), balancing performance with computational efficiency. Finally, the smaller models can be fine-tuned using synthetic data and expert feedback by measuring their accuracy and applicability. The combined cost of the model for data generation, model fine-tuning, and comprehensive evaluation for the work presented here was under $500, demonstrating that high-quality domain-specific models can be developed at lower costs and thus be accessible to institutions with more limited budgets (including academia and clinical settings). This contrasts favourably with multi-GPU or proprietary-API pipelines whose use for such a task can exceed several thousand dollars. This work has shown that the proposed workflow offers a financially sustainable path for institutions seeking to prototype or deploy medical AI systems without large capital investments.

Recent research in medical AI has concentrated on building and scaling generic foundation models [4, 37, 38] such as MONET [30], PANDERM [39], DERM1M [40], and pretrained models such as SkinGPT-4 [41]. These models have an emphasis on data annotation and short-context diagnostic tasks that fall short of critical needs in daily clinic such as long report generation.

From a clinical perspective, however, there is a growing need for decision support systems that can handle longer contexts, including detailed patient histories and comprehensive narratives [42, 43]. The development of DermatoLlama, a fine-turned 11-billion-parameter vision LLM, is capable of longer sequences and writing reports and allows for capturing more comprehensive clinical narratives in the output. DermatoLlama, with its 11B size and LoRA adapters reducing computing requirements, can be deployed on a system with a single Nvidia RTX4090 GPU, costing $\tilde{\$}$2,000 to set up.

With the LoRA adapters, multiple condition-specific adapters (e.g., rare dermatological diseases) were swapped in and out without duplicating the entire model. This model can also be used in cloud environments with GPUs, such as Google Colab with Nvidia A100 GPUs. A similar framework, LLaVa-Med, developed in the last few years [33] used 8×A100 GPUs for training, costing an estimated $400–600 for training alone and significant labour force. We integrated LoRA adapters to efficiently fine-tune our model with specialized medical functions while preserving high-quality report generation and complex instruction-following abilities. Our framework allows for privacy-preserving local refinement on proprietary data, allowing for full control of the data access. Leveraging synthetic data and knowledge transfer, we showed that clinically accessible-size models can achieve clinically relevant performance without requiring the need for massive architectures, thereby reducing dependence on sprawling LLMs and fully supervised training. Beyond diagnostic and treatment tasks targeting clinicians, this model is capable to adapt to everyday queries with its high quality writing capability - whether simplifying medical explanations for laypersons or providing supportive responses for stressed patients - underscoring its broader applicability beyond strictly clinical use cases. This feature of LLMs has potential to reduce burden of clinicians by improving the interpretability of clinical notes.

We compared our model with two commercial, closed-source vision–language models, GPT-4o (knowledge cut-off Oct/2023) and Gemini 2.0 Flash (knowledge cut-off Aug/2024), with the same queries as used to evaluate our model. As neither LLM has disclosed their training data, certain test items (especially in our Eval2), may already be included in their pre-training, thereby result in optimistic upper limits of performance rather than a perfectly fair comparison on unseen data (such as Eval1). In contrast, our model did not use any Eval1 or Eval2 data during training, yet it outperformed both commercial models on Eval1 and matches their accuracy on Eval2. The commercial models often produce very detailed, free-form answers, which results in lower surface-overlap metrics such as BLEU and ROUGE; by fine-tuning our model for concise, clinically targeted responses, we obtain higher scores on these metrics while better satisfying domain-specific requirements.

We demonstrate with AnnotatorMed how a purpose-built, free, open-source, and locally deployable tool can remove one of the largest bottlenecks in medical-AI development: secure, clinician-driven data curation. This tool is capable to run on local (e.g. hospital) infrastructure, behind a firewall, sensitive images or reports do not leave institutional storage, thereby satisfying privacy regulations while still enabling collaborative labelling projects. The hierarchical workflow proved particularly valuable in our use-case of dermatology, where a single finding often branches into multiple differential pathways; using the hierarchical design results in less screen clutter and increased annotation speed. In Eval1, clinicians answered 21,597 questions instead of the 72,000 that would have been required for 320 questions across 225 samples. Beyond labelling, the same interface can capture RLHF annotations or serve as a clinician test platform - allowing experts to rank, correct, or critique model outputs in context. Although we did not train an RLHF reward model in this study to iteratively improve AI performance, our architecture partitions feedback into versioned folders, facilitating iterative model alignment to allow for future improvements with

RLHF. The system requires some expert annotators, and the workload still scales linearly with dataset size, and transferring the hierarchy used here for dermatology to other specialties may require domain-specific adaptations. Nevertheless, by unifying multimodal input, flexible question logic, and secure on-premises deployment, AnnotatorMed provides a reusable back-end for high-quality data generation in clinical AI workflows.

We identified several challenges with our approach. Firstly, while synthetic data can increase the size of a dataset, it may fail to capture extremely subtle features such as those found in rare or advanced cases [44, 45]. Despite the low resource cost of training and deployment, there remains the time (and associated cost) required of clinical experts to validate and annotate large volumes of generated data. Our trained model contains 11B parameters and is therefore less resource intensive than trillion-parameter ones, however due to its size it may lack the high reasoning capacity required for further specific tasks. However, continued improvements in quantization (makes AI models smaller and faster), knowledge transfer, and incremental training will likely make large parameter models more tractable to be deployed as our model can now. As hardware and optimization techniques progress, we anticipate that models with 11B-parameters may be deployable on ordinary personal computers (without larger GPUs), making multi-agentic architectures viable for day-to-day clinical use. However, we emphasize that we have not yet implemented the extensive safeguards (external clinical validation, bias audits across diverse populations, adversarial stress-testing, and privacy controls) required for safe clinical deployment requiring formal regulatory review and rigorous prospective evaluation. Consequently, the model remains a research prototype and we do not recommend it be used to guide individual patient care.

Although our study primarily centres on a single-model architecture, the baseline approach of refining 10 billion-sized models with synthetic and local data can be applied to various medical fields, from radiology to pathology. Crucially, it demonstrates that trillion-parameter models are not mandatory for practical clinical utility. Moreover, this same approach provides a natural foundation for multi-agentic systems, should researchers choose to distribute specialized tasks—like image interpretation, textual reasoning, or patient triage—among smaller sub-models [46].

Looking ahead, our framework provides a strong basis for developing more advanced models that can be fine-tuned rapidly, allowing for iterative improvements in performance and metrics as clinical needs evolve. We anticipate that useful methods such as RLHF and retrieval-augmented generation (RAG, a technique that help AI retrieve reliable medical information) [47] will further mitigate hallucinations and improve contextual awareness [48], enabling 10 billion-sized models to extend their knowledge base in real time without significantly increasing parameter counts. In addition, creating baseline models that are domain-ready—and thus easily adaptable to local data—empowers individual institutions to tailor AI solutions to their specific patient populations while preserving privacy and cost-effectiveness. Finally, by incorporating complementary frameworks such as MONET [30], which offer automated concept annotation and auditing capabilities, these baseline models can achieve enhanced interpretability and clinical success. A hierarchical flow can be succesful by implementing new approaches for medicine (hierarchy-of-thought) analogous to the chain-of-thought experiments in LLMs [49].

In summary, by focusing on a single-model strategy and leveraging synthetic data for knowledge transfer, we demonstrate a framework that could be locally deployable in healthcare institutions to bypass the complexity of trillion-size LLMs and develop clinically applicable solutions on more modest hardware that can exceed current SOTA model performance and clinical text generation. This approach stands to accelerate the adoption of AI in healthcare by delivering a scalable, privacy-conscious, and efficient framework; one that not only supports current clinical needs but also points to a future in which specialized models can be further adapted for more diverse clinical

tasks and contexts.

Methods

Data

Overall Description

In the initial phase of our framework, SCALEMED focuses on data gathering, task structuring, and synthetic dataset creation. After merging the individual pipelines, we assembled a comprehensive dataset consisting of 112,695 image–text pairs from PubMed (as a high-trust data source), 89,628 image–text pairs from open-access repositories that provide only limited metadata (as a medium-trust data source), and synthetic instruction following data (as a low-trust data source) compromises 50,027 self-instruct prompts derived from 193 seed tasks, 72,400 knowledge-based tasks covering around 2,700 diseases generated from 40 seed tasks, and 9,963 multi-turn anamnesis interactions created from 137 disease-specific seed tasks. This integrated approach streamlines the workflow, enabling researchers to progress seamlessly from raw data sources to a carefully curated synthetic dataset, which serves as the foundation for subsequent model training and evaluation. Data generation details are shown in Fig. 4. No human participants or personal identifying data were involved in this study. All image–text pairs, prompts, and derived synthetic data originate from publicly available open-access repositories and literature, thereby obviating the need for additional ethical approval. All licenses (only commercially available sources) and usage rights were verified to ensure that no patient privacy or confidential information was compromised. The dataset specification are shown in Table 2.

Data Sources and Preprocessing: Open Access PubMed Articles

We utilized the pre-existing BIOMEDICA dataset [27], which comprises image–caption pairs and associated metadata extracted from millions of open-access biomedical articles available on PubMed (PM). We filtered this dataset to identify figures relevant to dermatology using 'Skin Lesion' keyword and selected samples with permissive licenses allowing research and derivative works (e.g., CC0, CC BY, CC BY-SA, CC BY-ND). From this filtered subset, we extracted 56,258 dermatology-related images after excluding non-skin images such as graphical figures. For each image, we paired it with its associated textual context, primarily figure captions and relevant inline mentions identified within the BIOMEDICA metadata [27]. This formed the basis for generating image-grounded question-answer pairs. We also developed a PubMed Central scrapping tool to collect data from any field from the latest PubMed database.

Data Collection and Preprocessing: Open-Access Datasets

We identified open-access datasets relevant to our target domain, dermatology, prioritizing collections with clear licensing terms and comprehensive documentation. For dermatology, this often included image repositories that covered various skin conditions and age groups, thereby enhancing model generalizability. Where necessary, images were resized or converted to consistent resolutions, and any mislabelled or corrupted files were removed. Pertinent metadata—such as lesion type, imaging modality or diagnostic tags—was extracted and centralized for easy reference. This

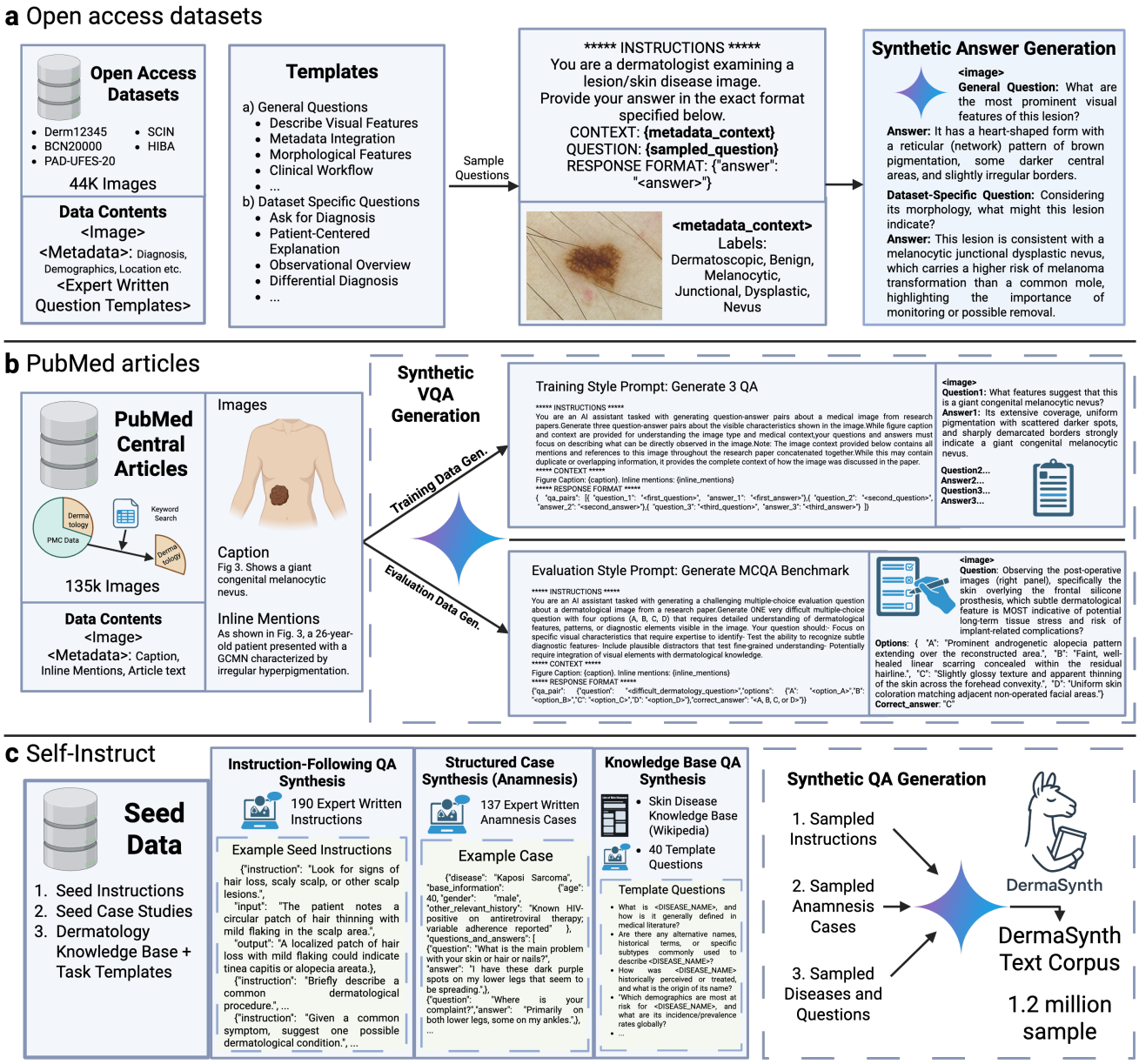

Figure 4: **Overview of synthetic data generation workflow. a,** 44k open access dermatology images with metadata were used to generate grounded QA pairs (medium-trust/hallucination). **b,** 37k open access PubMed images with captions and inline mentions were used to generate descriptive QA for training and MCQs for evaluation (high-trust, low-hallucination). **c,** 1,014k Synthetic instruction-following cases were generated based on self-instruct, knowledge base and anamnesis to cover clinical reasoning (low trust, high-hallucination).

coordinated preprocessing effort produced a unified, high-quality dataset, suitable for both human annotation and automated synthetics generation within the SCALEMED framework.

In this study, we used DERM12345 [26], BCN20000 [50], HIBA [51], PAD-UFES-20 [52] and SCIN [8] datasets which all of them have commercial licenses to be used with SOTA models. These repositories collectively comprised 13,444 clinical images (30%) and 31,370 dermatoscopic images (70%). Some were accessed via the isic-cli library, while others required downloads from institutional or project-specific websites.

Synthetic Data Generation

We employed multiple pipelines to generate a large and diverse synthetic dataset (DermaSynth) targeting various aspects of dermatological knowledge and reasoning. The generation process primarily utilized Google's Gemini models (Gemini 2.0 Flash, date: from November 2024 to January 2025) via API calls. Gemini 2.0 Flash can generate high quality synthetic data even for limited basic seed tasks.

VQA Generation from Image Datasets (Open Access & Biomedica)

- **Open Access Datasets:** For each image in the open-access datasets (such as DERM12345), we first converted its structured metadata into a concise natural language string (e.g., "Metadata - Label: Melanoma, Age: 45, Sex: Female, Location: Forearm"). We then used predefined prompt templates containing both "general" (e.g., "Describe the visual features in the image") and "dataset-specific" (e.g., "What is the diagnosis for this lesion located on the localization?") question types. For each image, one general and one dataset-specific questions were randomly selected, filled with metadata information, and combined with the metadata string and image as input to Gemini. We requested two synthetic QA pairs per image, formatted in JSON for easy parsing. For the open-access datasets, 92,020 samples were generated, and 2,392 were excluded because the data could not be parsed.

- **BIOMEDICA Dataset:** For each selected BIOMEDICA sample, we concatenated its caption and associated context mentions into a single text blob. This text, along with the image, was provided to Gemini. Gemini rejected 14,429 images because of its guard system. The rest of 41,829 images were splitted into training and evaluation splits. We used two prompt styles: a "training style" prompt requesting three descriptive QA pairs about visible characteristics based on the context, and an "evaluation style" prompt (used for generating Eval2) requesting either a multiple-choice question or a challenging open-ended question based on the image and context.

Instruction Based Generation

We utilized self-instruction techniques to generate diverse text-based and conversational data, starting from small sets of human-curated seed tasks. All seed tasks were generated by ChatGPT o1 pro and edited by expert clinicians to accelerate the seed task generation and to create seed tasks from different perspectives such as 'explain like I am 5'. Our framework is flexible enough to integrate with various types of seed tasks for diverse use cases, and we have designed it to be as comprehensive as possible.

1. Instruction-Following QA Synthesis: Starting with 193 seed instruction-response pairs (18 classification, 175 non-classification, validated by clinicians Gulsum Gencoglan (G.G.), Ece Gokyayla (E.G.), Ozan Erdem (O.E.)), we iteratively prompted Gemini. In each round, a mix of two human- and two model-generated examples were provided as context. The model was asked to generate new instructions and labelled them as classification or non-classification tasks. For classification tasks, it generated class labels and inputs; for non-classification, it generated input-output pairs. Outputs were parsed using regular expressions, and near-duplicates were filtered based on semantic similarity. This pipeline generated approximately 50,027 instruction-following samples.

2. Knowledge-Base (KB) Task Synthesis: We constructed a knowledge base covering approximately 2,700 dermatological diseases [53]. The knowledge base was sourced partially from Wikipedia's list of skin conditions using Wikipedia-API. We defined 40 seed task templates designed by clinicians (e.g., "List key triggers for <DISEASE_NAME>", "Describe the typical presentation of <DISEASE_NAME>"). For each disease, we instantiated relevant templates and content titles in each disease page, provided disease information as context (extracted from the KB), and prompted Gemini to answer authoritatively without directly referencing the provided context. This generated approximately 72,400 QA pairs grounded in dermatological knowledge.

3. Structured Case Synthesis (Anamnesis): We used 137 seed dermatological case studies based on the most common 137 dermatological disease (formatted as JSON, validated by clinicians) reflecting common conditions and doctor-patient interactions based on a standardized clinical questionnaire (17 sections, 104 questions) [35]. We prompted Gemini by providing a few example cases and asking it to generate a new case study following a strict JSON schema (including sections like base_information, questions_and_answers, final_diagnosis, treatment_of_choice). Robust parsing handled variations in model output formatting. This generated 892,814 samples for 9,963 synthetic multi-turn case interactions.

Final DermaSynth Dataset

The outputs from these pipelines were aggregated to form the DermaSynth dataset, comprising approximately 1.2M synthetic samples. All experiments were designed to remain within a modest budget. End-to-end evaluation of the state-of-the-art model was run on OpenAI API; yielding an inference cost of $100 for GPT-4o testing. Synthetic-data creation was performed with the Gemini API: Around 400k prompt–completion cycles were issued in batched mode. At the current rate of $0.55 per million output tokens, total generation cost was around $200. No additional charges were incurred for AnnotatorMed, which was hosted on a free version of Firebase. Thus, the complete pipeline—data generation, fine-tuning, and evaluation—was executed for well under $500, demonstrating that high-quality domain models can be developed on a budget accessible to most academic or clinical research groups.

Evaluation Data

We created three distinct evaluation benchmarks to assess model performance comprehensively.

1. Eval1 (Expert-Curated Dermoscopic Image MCQA): Expert clinicians (G.G., E.G., O.E.) designed 320 hierarchical questions (single choice, multi choice, open-ended) based on the

Data Source / Task	# Samples	Notes
PubMed image–text pairs	112,695	37,565 images with three QA pairs
Open Access Repositories	89,628	44,814 images with two QA pairs
DERM12345 (Train)	9,860	Dermoscopic
DERM12345 (Test)	2,485	Dermoscopic
BCN20000 (Train)	11,761	Dermoscopic
BCN20000 (Test)	6,075	Dermoscopic
PAD-UFES-20	2,292	Clinical
SCIN	10,858	Clinical
HIBA	1,572	1,270 Clinical, 302 Dermoscopic
Subtotal	44,814	13,444 Clinical, 31,370 Dermoscopic
Self-Instruct	50,027	193 seed task
Knowledge-based tasks	72,400	40 seed task, based on Wikipedia [53]
Multi-turn anamnesis	892,814	9,963 multi-turn case and 137 seed task
Eval1	21,597	225 samples of DERM12345 - Test
Eval2	4,264	PubMed test split
Eval3	1,998	LLM-generated questions, Gemini used
Total	1,245,423	

Table 2: Characteristics of DermaSynth dataset.

two-step pattern analysis in lesion diagnosis. They then annotated 225 images from the test split of the DERM12345 dataset (approximately 5 images per class across 40 classes) using these questions, creating a challenging open-ended VQA benchmark that requires nuanced clinical reasoning. In total, 93 main questions and 227 subquestions were developed, organized into five main sections—intro, clinical, dermoscopy-basic, dermoscopy-melanocytic, dermoscopy-nonmelanocytic. The dermoscopic section follows a two-step pattern analysis [29]: Step 1 focuses on basic lesion context, while Step 2 involves determining whether the lesion is melanocytic or non-melanocytic, further subdivided into nine detailed levels. The summarization section likewise employs two steps: (1) evaluating the lesion's global and local patterns to classify it as a benign nevus, suspect melanocytic lesion, or melanoma, and (2) answering final diagnosis decision questions. Unrelated questions were omitted from annotation of Eval1 set, such as certain clinical inquiries that did not apply for dermoscopic images. Additionally, each section features two special questions prompting a summary report and asking for the model's confidence level and rationale for SOTA models. To generate these annotations efficiently, we first used GPT-4o model to produce preliminary outputs, then applied multiple data validation steps to align the synthetic data with our backend and UI structures. These checks included many steps such as removing invalid answers, correcting mislabelling (e.g., "Brown globules" → "globules"), verifying IDs, and ensuring that open-text fields, missing IDs, and option IDs matched the required format. We also initiated a new thread in the OpenAI API for each image and split the questions appropriately to handle them consistently. This approach ensured that the synthetic annotations not only conformed to our data structures but also upheld the quality needed for reliable evaluation.

2. Eval2 (Clinical Image MCQA): This benchmark consists of multiple-choice questions generated using the "evaluation style" prompt applied to the test split of BIOMEDICA image-text data. These questions test comprehension and reasoning based on clinical images and their

associated context (captions, inline mentions).

3. Eval3 (LLM-Generated Text-Only MCQA): Generated using the "Multiple-choice / balanced benchmark" pipeline, this benchmark comprises text-only multiple-choice questions covering various aspects of dermatology (such as diagnosis, treatment, histopathology) balanced across difficulty levels (beginner, mid, expert). It assesses the model's textual knowledge base.

Model Training and Evaluation

In this section, we discuss the model training and evaluation strategies employed to ensure robust performance across different use cases, with flexibility for both vision-based and text-based tasks. Depending on the specific needs of the application and the medical domain, either vision or text-based LLMs can be utilized. Multi-modal models, which combine both vision, text, and other data processing capabilities, offer additional versatility for applications that require integrated analysis of both image and textual data. Text-based LLMs are particularly effective for tasks involving structured data, such as medical record analysis, diagnostic reasoning, or clinical decision support, while vision-based models excel in image classification and visual diagnostic applications. The choice between these model types depends on the nature of the data and the specific requirements of the medical field, ensuring the system's adaptability to a variety of contexts within healthcare. Finally, different prompts were used for evaluation of classification and text metrics. For classification metrics, single and multi-choice questions were used, and all question types were used to calculate text metrics.

Teacher–Student Paradigm and Model Training

To balance performance and computational efficiency, we adopted a teacher student approach. First, we used a larger teacher model, Gemini 2.0 Flash, to guide a lighter multimodal student model, Llama-3.2-11B-Vision in 16-bit (developed by Meta), distilling key insights via LoRA fine tuning [10]. This process reduces hardware overhead while retaining high level reasoning capabilities. The DermatoLlama models were trained and evaluated in Python (v3.10.13) using LLaMa-Factory (v0.9.2) [54] and DeepSpeed (v0.16.3). Training was performed on a readily attainable workstation equipped with four Nvidia RTX 4090 GPUs (96 GB combined VRAM) and 512 GB of system RAM. For inference and benchmarking, we used a standard desktop containing a single RTX 4090 GPU (24 GB VRAM) and 96 GB RAM, confirming that the model can be deployed on a single-GPU setup.

Low–Rank Adaptation (LoRA) Fine-Tuning

LLMs normally require updating all weights during fine-tuning, a process that is memory-intensive and slow. LoRA [10] instead learns only a small, low-rank update to each weight matrix while the original (pre-trained) weights stay fixed.

For any linear layer with pretrained weights $\mathbf{W}_0$, LoRA adds a learnable correction

$$\boxed{\mathbf{W} \;=\; \mathbf{W}_0 + \mathbf{B}\mathbf{A}^{\mathsf{T}}}$$

where $\mathbf{A} \in \mathbb{R}^{k \times r}$ and $\mathbf{B} \in \mathbb{R}^{d \times r}$ have a very small *rank* r.

In practice, we applied LoRA to DermatoLlama by choosing a rank of 16, dropout value of 0.05, and input/output cut-off of 2,048 for targeted modules. Only the small LoRA factors (the A and B matrices) were updated during fine-tuning, while all pretrained weights remained frozen. After training, we merged the learned low-rank updates back into the original weight matrices, so the deployed model incurs no extra memory or latency overhead. This strategy achieved state-of-the-art dermatology classification while fitting comfortably on a four RTX 4090 GPU.

After training, the low-rank product was merged into $\mathbf{W}_0$ for zero-overhead inference or kept separate so the same base model serves multiple domain-specific LoRA adapters with minimal storage.

Model Auditing

To confirm model improvements, we incrementally introduced 10k, 50k, 100k, 200k, and Full samples from our synthetic dataset (DermaSynth), measuring performance at each stage. Test split of DERM12345 were excluded during training. Performance was evaluated at each stage using the evaluation datasets. Accuracy, BLEU-4, ROUGE-1, ROUGE-2, and ROUGE-L scores were measured to assess the performance of DermatoLlama and SOTA models.

Tools

We developed AnnotatorMed, an open-source, lightweight annotation platform written in Flutter (v3.22.3) and deployable both on the web and mobile devices. A Firebase backend manages labelling records although any database can be substituted for on-premises installations. In this study, we also developed a PubMed scrapping tool to filter any specific field from PubMed Central database.

The graphical interface is driven by a bespoke, hierarchical data model—question → option → sub-question → sub-option—that condenses nested clinical logic into a single screen flow. This structure supports three question formats (single choice, multiple choice, open-ended); when an annotator selects an option, only the relevant follow-up items are revealed, reducing click overhead and cognitive load. The same component library can be switched to a flat, non-hierarchical mode if desired. All annotations are exportable as CSV or JSON for direct ingestion into downstream fine-tuning or evaluation pipelines. Using this interface, three dermatologists (G.G., E.G., O.E.) labeled 225 images from the DERM12345 test split. The expand-on-demand UI decreased annotation time while ensuring that every decision—clinical finding, dermoscopic cue, or free-text comment—was stored in a consistent, machine-readable format.

License Issues

All synthetic data derived from PubMed articles and open-access datasets that permit commercial usage; however, the generation relied on OpenAI and Gemini APIs, each with their own usage restrictions. The VLM model, DermatoLllama, was fine-tuned using Llama-3.2-11B-Vision model. Consequently, DermaSynth and DermatoLlama are released under CC-BY-NC 4.0, prohibiting direct commercial exploitation. This aligns with:

- Llama's Non-Commercial License: Derivative projects inherit non-commercial constraints.
- Upstream Model Limitations: Some instruction data stems from providers prohibiting commercial uses.

- Safety and Maturity Concerns: Additional safeguards are required before broad clinical deployment.

  Researchers must independently verify compliance with all upstream licenses and consider implementing safety measures before adopting these models in real-world patient-care contexts.

Data Availability

BIOMEDICA dataset is publicly available from `https://huggingface.co/BIOMEDICA`. Open access datasets are all publicly available and can be accessed from: DERM12345 (`https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DAXZ7P`), BCN20000 (`https://figshare.com/articles/journal_contribution/BCN20000_Dermoscopic_Lesions_in_the_Wild/24140028/1`), SCIN (`https://github.com/google-research-datasets/scin?tab=readme-ov-file`), HIBA (`https://api.isic-archive.com/collections/175/`), and PAD-UFES-20 (`https://data.mendeley.com/datasets/zr7vgbcyr2/1`). DermaSynth dataset and DermatoLlama models are available at `https://huggingface.co/DermaVLM`.

Code Availability

Codes are available at `https://github.com/DermaVLM`.

Acknowledgement

Abdurrahim Yilmaz has been funded by the President's PhD Scholarships at Imperial College London. Donghee Choi and Joram M. Posma are supported by the Horizon Europe project CoDiet. The CoDiet project is funded by the European Union under Horizon Europe grant number 101084642. CoDiet research activities taking place at Imperial College London is supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant number 101084642). All figures were created with BioRender.

Competing Interest

The authors declare no competing interests.

References

[1] Singhal, K. *et al.* Toward expert-level medical question answering with large language models. *Nature Medicine* 1–8 (2025).

[2] Mamede, S. & Schmidt, H. G. Making large language models into reliable physician assistants. *Nature Medicine* 1–2 (2025).

[3] Goh, E. *et al.* Gpt-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nature Medicine* 1–6 (2025).

[4] Liu, X. *et al.* A generalist medical language model for disease diagnosis assistance. *Nature Medicine* 1–11 (2025).

[5] Lu, S. *et al.* General lightweight framework for vision foundation model supporting multi-task and multi-center medical image analysis. *Nature Communications* **16**, 2097 (2025).

[6] Tanno, R. *et al.* Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine* **31**, 599–608 (2025).

[7] Jiang, Y. *et al.* Evaluating general vision-language models for clinical medicine. *medRxiv* 2024–04 (2024).

[8] Ward, A. *et al.* Creating an empirical dermatology dataset through crowdsourcing with web search advertisements. *JAMA Network Open* **7**, e2446615–e2446615 (2024).

[9] Ahmed, M. I. *et al.* A systematic review of the barriers to the implementation of artificial intelligence in healthcare. *Cureus* **15** (2023).

[10] Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models. *ICLR* **1**, 3 (2022).

[11] Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems* **36**, 10088–10115 (2023).

[12] Kiyasseh, D., Cohen, A., Jiang, C. & Altieri, N. A framework for evaluating clinical artificial intelligence systems without ground-truth annotations. *Nature Communications* **15**, 1808 (2024).

[13] Zou, J. & Topol, E. J. The rise of agentic ai teammates in medicine. *The Lancet* **405**, 457 (2025).

[14] Hanser, T. *et al.* Data-driven federated learning in drug discovery with knowledge distillation. *Nature Machine Intelligence* 1–14 (2025).

[15] Chen, X. *et al.* Evaluating large language models and agents in healthcare: key challenges in clinical applications. *Intelligent Medicine* (2025).

[16] Li, B. *et al.* Mmedagent: Learning to use medical tools with multi-modal agent. *arXiv preprint arXiv:2407.02483* (2024).

[17] Fallahpour, A., Ma, J., Munim, A., Lyu, H. & Wang, B. Medrax: Medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673* (2025).

[18] Wang, Z., Wu, J., Low, C. H. & Jin, Y. Medagent-pro: Towards multi-modal evidence-based medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968* (2025).

[19] Nori, H. *et al.* Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452* (2023).

[20] Wang, Y. *et al.* Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560* (2022).

[21] Peng, B., Li, C., He, P., Galley, M. & Gao, J. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277* (2023).

[22] Zhang, S. *et al.* Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792* (2023).

[23] meta-llama (Meta Llama) — huggingface.co. `https://huggingface.co/meta-llama`. [Accessed 06-05-2025].

[24] Utti, V., Bikia, V., Agarwal, A. A. & Daneshjou, R. Integrating artificial intelligence in dermatological cancer screening and diagnosis: Efficacy, challenges, and future directions. *Annual Review of Biomedical Data Science* **8** (2025).

[25] Movva, R., Koh, P. W. & Pierson, E. Using unlabeled data to enhance fairness of medical ai. *nature medicine* **30**, 944–945 (2024).

[26] Yilmaz, A., Yasar, S. P., Gencoglan, G. & Temelkuran, B. Derm12345: A large, multisource dermatoscopic skin lesion dataset with 40 subclasses. *Scientific Data* **11**, 1302 (2024).

[27] Lozano, A. *et al.* Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. *arXiv preprint arXiv:2501.07171* (2025).

[28] Team, G. *et al.* Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[29] Argenziano, G. *et al.* Dermoscopy of pigmented skin lesions: results of a consensus meeting via the internet. *Journal of the American Academy of Dermatology* **48**, 679–693 (2003).

[30] Kim, C. *et al.* Transparent medical image ai via an image–text foundation model grounded in medical literature. *Nature Medicine* **30**, 1154–1165 (2024).

[31] Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *Advances in neural information processing systems* **35**, 27730–27744 (2022).

[32] Tu, T. *et al.* Towards conversational diagnostic artificial intelligence. *Nature* 1–9 (2025).

[33] Li, C. *et al.* Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* **36**, 28541–28564 (2023).

[34] Cirone, K., Akrout, M., Abid, L. & Oakley, A. Assessing the utility of multimodal large language models (gpt-4 vision and large language and vision assistant) in identifying melanoma across different skin tones. *JMIR dermatology* **7**, e55508 (2024).

[35] Johri, S. *et al.* An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine* 1–10 (2025).

[36] Zhang, G. *et al.* Leveraging long context in retrieval augmented language models for medical question answering. *npj Digital Medicine* **8**, 239 (2025).

[37] Kiraly, A. P. *et al.* Health ai developer foundations. *arXiv preprint arXiv:2411.15128* (2024).

[38] Tu, T. *et al.* Towards generalist biomedical ai. *Nejm Ai* **1**, AIoa2300138 (2024).

[39] Yan, S. *et al.* A general-purpose multimodal foundation model for dermatology. *arXiv preprint arXiv:2410.15038* (2024).

[40] Yan, S. *et al.* Derm1m: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. *arXiv preprint arXiv:2503.14911* (2025).

[41] Zhou, J. *et al.* Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4. *Nature Communications* **15**, 5649 (2024).

[42] Gui, H., Omiye, J. A., Chang, C. T. & Daneshjou, R. The promises and perils of foundation models in dermatology. *Journal of Investigative Dermatology* (2024).

[43] Zarfati, M. *et al.* Exploring the role of large language models in melanoma: A systematic review. *Journal of Clinical Medicine* **13**, 7480 (2024).

[44] Omar, M. *et al.* Sociodemographic biases in medical decision making by large language models. *Nature Medicine* 1–9 (2025).

[45] Alber, D. A. *et al.* Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine* 1–9 (2025).

[46] Kim, H. *et al.* Small language models learn enhanced reasoning skills from medical textbooks. *npj Digital Medicine* **8**, 240 (2025).

[47] Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* **33**, 9459–9474 (2020).

[48] Chang, C. T. *et al.* Red teaming chatgpt in medicine to yield real-world insights on model behavior. *npj Digital Medicine* **8**, 149 (2025).

[49] Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **35**, 24824–24837 (2022).

[50] Hernández-Pérez, C. *et al.* Bcn20000: Dermoscopic lesions in the wild. *Scientific data* **11**, 641 (2024).

[51] Ricci Lara, M. A. *et al.* A dataset of skin lesion images collected in argentina for the evaluation of ai tools in this population. *Scientific Data* **10**, 712 (2023).

[52] Pacheco, A. G. *et al.* Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in brief* **32**, 106221 (2020).

[53] Wikipedia - list of skin conditions. https://en.wikipedia.org/wiki/List_of_skin_conditions. [Accessed 15-01-2025].

[54] Zheng, Y. *et al.* Llamafactory: Unified efficient fine-tuning of 100+ language models (2024). URL http://arxiv.org/abs/2403.13372.