DermIQ-VLM: A Methodological Approach for Dermatological Reasoning in Vision Language Models under Low-Resource Constraints

Ismam Nur Swapnil[a,1], Tanvir Ahmed Khan[a,1], Aranya Saha[a,1], Mohammad Ariful Haque[a]

[a] *Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh*

Abstract

Vision-Language Models (VLMs) show promise in medical image analysis, yet their capacity for structured reasoning in complex domains like dermatology is often limited by data scarcity and the high computational cost of advanced training techniques. To address these challenges, we introduce **DermIQ-VLM**, a VLM developed through a multi-stage, resource-efficient methodology designed to emulate a dermatologist's diagnostic process. Our primary contribution is Adaptive Grouped Relative Policy Optimization (AGRPO), a hybrid curriculum-RL technique that stabilizes the powerful but data-intensive Grouped Relative Policy Optimization (GRPO) framework. By incorporating probabilistic ground-truth hints during training, AGRPO ensures a consistent learning signal even with a minimal number of candidate generations (2–3), enabling convergence in low-resource settings. The training pipeline first employs AGRPO for reasoning-oriented disease recognition, followed by supervised fine-tuning for conversational ability. To mitigate factual errors introduced during this step, we then align the model using Direct Preference Optimization (DPO), leveraging a Knowledge Graph-based system as a scalable proxy for expert preference. A preliminary evaluation on a curated dermatological dataset demonstrates that our proposed methodology yields notable performance gains over standard fine-tuning approaches. These findings validate the potential of our pipeline as a feasible pathway for developing specialized, reliable VLMs in resource-constrained environments.

Keywords:
Vision Language Model, Low-Resource, GRPO, AGRPO, DPO, DermIQ-VLM

---

1. Introduction

Thousands of skin lesions are annually assessed around the world, leading to a significant increase in the need for dermatological consultations. As a result, the demand for efficient, accurate, and reliable AI support in dermatology is becoming increasingly critical to meet this growing challenge. Recently, in the field of medical Visual Question Answering (VQA) [1] , Vision-Language Models like GPT-4o [3] and Grok [4] are available to support clinicians and patients in the interpretation of skin lesion images, showing promising reasoning abilities. It is important for transparent decision making in dermatology; both the clinician and the patient must be aware of not only the final diagnosis, but also the rationale behind it. However, current medical VLMs provide only the final

---

[1] These authors contributed equally to this work.

diagnosis or provide a very shallow explanation based on memorized pattern [5], showing neither comprehensible reasoning nor genuine step-by-step reasoning. This void in interpretability and reliability presents significant challenges for clinical uptake of these technologies.

The limited reasoning ability of existing dermatological VLMs is largely due to the shortcoming of Supervised Fine-Tuning (SFT), which is a common paradigm for adapting general models to medical tasks [8]. SFT often leads to overfitting and shortcut learning [7] which affects the performance on unseen dermatological data, and does not encourage the development of models having explicit reasoning capabilities since their emphasis is only upon final diagnoses. Reasoning in AI models, particularly for complex tasks like dermatological imaging, often benefits from reinforcement learning with human feedback (RLHF). However, RLHF is memory-intensive, posing challenges for scalability. Alternatively, fine-tuning with custom chain-of-thought (CoT) [21] datasets can be effective but requires creating high-quality datasets, which demands significant resources and is difficult to scale. Methods like Group Relative Policy Optimization (GRPO) [37], a reinforcement learning approach, are more appealing as they enhance reasoning with lower computational requirements. Despite their potential, GRPO's application in medical visual question answering remains largely unexplored. To address this, we propose a novel GRPO variant that further optimizes memory efficiency, enabling more scalable and effective reasoning for dermatological imaging applications while maintaining high performance.

A key challenge with current Medical Vision-Language Models (VLMs) is their tendency to produce hallucinated or factually inaccurate responses [15]. Most existing approaches struggle to balance clear reasoning with factual accuracy, which restricts their utility in medical applications. To overcome these limitations, we employ Knowledge Graph Retrieval Augmented Generation (KG-RAG) [14], which enhances VLMs by incorporating domain-specific knowledge. KG-RAG leverages structured knowledge graphs—databases of entity relationships that store information about entities like skin conditions, symptoms, causes, or treatments—to enrich model responses. When users input dermatological images or questions, the system identifies relevant clinical concepts and retrieves corresponding facts from the knowledge graph, integrating them into its reasoning process. This approach ensures responses are both medically accurate and contextually relevant, significantly reducing hallucinations. As a result, KG-RAG enhances the depth and reliability of answers, making it a critical tool for advancing AI in patient care.

However, despite KG-RAG's ability to provide robust factual grounding, it does not guarantee that the model fully internalizes these responses. This limitation hinders the complete realization of KG-RAG's potential. To address this issue, Direct Preference Optimization (DPO) [29], a reinforcement learning-based method, can fine-tune the decision-making of the model to internalize the facts offered by KG-RAG. DPO uses two responses for each input, one based on the knowledge graph and one without the knowledge graph, and marks the knowledge-based response as preferred and the other as rejected. In this way, it shifts the model's tendency towards responses that depend on actual facts, emphasizing better accuracy and detail through a preference-based mode of training. With time, the approach will improve the model's response quality without needing continuous outside guidance. In dermatology, DPO works hand in hand with KG-RAG to ensure that the model is accessing trustworthy information as well as internalizing it efficiently to use this knowledge appropriately for clinical application.

In this paper, we present **DermIQ-VLM**, a dermatology-specific vision-language model that integrates structured reasoning, domain knowledge to support trustworthy diagnostic decision-making. Specifically, our contributions are as follows:

- **Dataset Curation:** We curated a comprehensive dermatological VQA dataset by collecting and consolidating image-question-answer pairs from trusted medical sources. This dataset served as the foundation for training, fine-tuning, and evaluating all stages of our model development.
- **Reasoning-Oriented Optimization:** We applied *Adaptive Grouped Relative Policy Optimization (AGRPO)* using the curated dataset to guide the model toward structured visual reasoning. This encouraged the model to associate dermatological features with diagnostic decision-making patterns typically used by clinicians.
- **Low-Resource Training:** We trained our vision-language model under constraints of limited response generations, reducing computational overhead. This setup enabled faster convergence while maintaining model performance, making the approach suitable for low-resource settings.
- **Clinical Alignment via Expert Feedback:** We used *KG-RAG* with the model to generate responses, which we labeled as expert responses. These were paired with non-KG-RAG model outputs to create a preference dataset. We then applied *DPO* to train the model using this dataset, improving factual accuracy even without KG-RAG at inference time.

DermIQ-VLM presents a significant step towards robust, transparent, and reliable AI in dermatology, demonstrating superior generalization and performance efficiency, thereby offering substantial practical utility for clinical adoption in resource-constrained settings.

2. Related Works

Vision-Language Models (VLMs) are machine learning models that integrate visual and textual data to perform tasks requiring understanding and reasoning across both modalities. They are designed to process images and text jointly, enabling applications like image captioning, visual question answering, and multimodal reasoning. VLMs typically consist of two primary components that work together to process and fuse visual and textual information: Image Encoder and Text Decoder. There is also some Fusion Mechanism to transfer modalities.

Large language models (LLMs) and foundation models are increasingly integral to clinical artificial intelligence workflows, supporting tasks such as answering medical questions and drafting patient reports [34]. These models combine deep learning's pattern-recognition capabilities with LLMs' contextual understanding to automate diagnostic processes, such as interpreting medical images and generating clinician-like explanations [34]. Specialized systems like ChatCAD and OphthUS-GPT integrate LLMs into computer-aided diagnosis workflows, transforming visual outputs into natural language reports [36]. However, pioneering studies applying multimodal LLMs in areas like ECG analysis or medical image reporting are often limited to single modalities, such as ECG or chest X-rays, highlighting a gap in comprehensive multimodal integration [34, 35].

The integration of vision-language models (VLMs) has significantly advanced medical AI, particularly in fields like dermatology, radiology [33], pathology [11], and broader clinical applications [22, 13]. VLMs such as those explored in recent studies [23, 10, 19, 16] demonstrate robust multimodal interpretability, enabling the processing of both visual and textual medical data. However, standard fine-tuning (SFT) often results in superficial learning, which is insufficient for complex clinical diagnostics [20]. To overcome this, advanced training strategies incorporate multistage approaches to foster deeper reasoning capabilities.

Interpretability remains a cornerstone for clinical acceptance of AI models. Techniques like Chain of Thought (CoT) prompting [21] promote sequential reasoning in language models [18], but the clinical validity and justification of these explanations are critical [12]. Fine-tuning on structured, domain-specific CoT data enables the generation of clinically coherent and grounded reasoning narratives, enhancing the reliability and transparency of diagnostic outputs in medical AI applications.

In the field of medicine, where truth is paramount, generative models often struggle with hallucination, producing outputs that may not align with verified knowledge [28]. To address this, Retrieval-Augmented Generation (RAG) [25] has emerged as a critical approach, particularly Knowledge Graph-based RAG (KG-RAG) [26], which leverages structured knowledge representations to enhance factual reliability. This method is especially valuable in complex domains like medicine, where structured knowledge bases can anchor model outputs to verified information, reducing errors [27].

Advanced alignment techniques, such as Reinforcement Learning (RL), have further refined model performance. Building on Reinforcement Learning from Human Feedback (RLHF) [31], Direct Preference Optimization (DPO) [29] optimizes models using preference data to prioritize accurate and contextually relevant outputs. Recent advancements include methods like Group Relative Policy Optimization (GRPO) [32], which supports structured reasoning, often followed by DPO to enhance factual accuracy when integrated with KG-RAG. In dermatology, prior AI research has primarily focused on image classification, achieving high accuracy in lesion identification using large datasets [6, 17, 2]. However, these classifiers typically lack interactivity and fail to provide detailed rationales for their outputs. In contrast, modern vision-language models (VLMs) designed for medical applications emphasize diagnostic reasoning within conversational frameworks. These models deliver interpretable, step-by-step rationales, integrating visual and medical knowledge to enhance both accuracy and user interaction.

3. Proposed Training Methodology

Our training framework follows a structured development process. At stage-1, we begin with AGRPO-based reinforcement learning to initialize visual disease detection capabilities. Later at stage-2, Supervised fine-tuning further enhances multi-turn conversational performance.Finally, To improve factual accuracy and reduce hallucinations, at stage-3, we align the model using Direct Preference Optimization (DPO), leveraging a preference dataset. The complete training methodology is summarized in Figure 1.

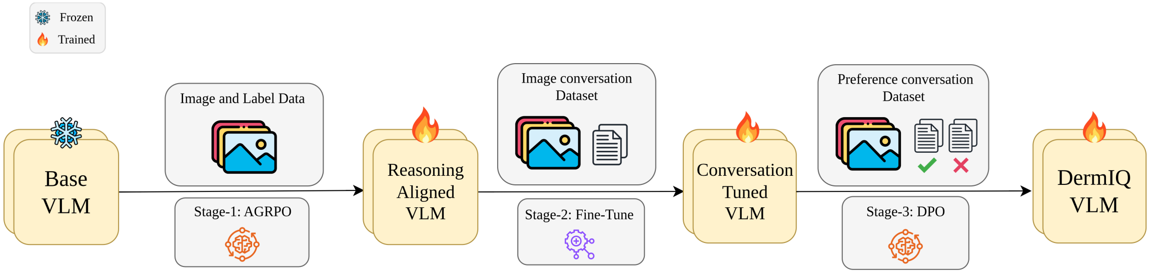

Figure 1: Proposed Training Methodology

3.1. Dermatological Dataset Creation

3.1.1. Image and Label Dataset for AGRPO:

We have developed a specialized dermatological Visual Question Answering (VQA) dataset, specifically designed to address the challenging task of distinguishing among seven skin diseases known for their often similar appearances. These are: *Dermatitis*, *Basal Cell Carcinoma*, *Rosacea*, *Psoriasis*, *Actinic Keratosis*, *Seborrheic Keratosis*, and *Melanoma*. The images for these diseases were sourced from a renowned dermatology dataset—the **Dermnet** dataset. We selected 700 high-quality images, with 100 images retained for each of the seven diseases. Each image is accompanied by a carefully designed Question & Answer (Q&A) pair which can also be called as image label. For example, a Question & Answer (Q&A) pair is presented as follows:

User Prompt: What is the name of the disease?
Ground Truth: Seborrheic Keratosis

3.1.2. Image and Conversation Dataset for Fine-tuning:

Our Supervised Finetuning (SFT) dataset is carefully developed to train the model not only to recognize skin diseases but also to explain its diagnostic reasoning and respond to related questions in a clear, human-understandable way. Each data instance in this dataset consists of a dermatological image, a user question, and a detailed, multi-part ground-truth answer. This answer reflects the step-by-step thought process of a practicing dermatologist by methodically examining visual features—such as color, texture, and shape—and combining these observations to arrive at a logical diagnosis. The example below demonstrates the typical structure of a data point in the SFT dataset:

Example-1: Cause-Based Q&A

User Prompt: What causes Seborrheic keratosis?

The exact cause of Seborrheic keratoses is not fully understood, but they tend to increase with age and may be influenced by factors such as genetics, sunburn, dermatitis, and skin friction. Mutations in specific genes—such as *FGFR3*, *PIK3CA*, and *EGFR*—have been associated with their development.

Ultraviolet (UV) exposure may also contribute to their formation. In some cases, the use of epidermal growth factor receptor (EGFR) inhibitors has been linked to an increased number of these lesions.

Example-2: Image Description Q&A

User Prompt: Describe the image.

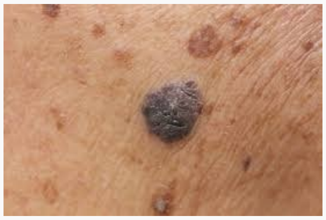

Figure 2: Seborrheic Keratosis

Color and Pigmentation: The lesion appears light brown to grayish-brown in color. There are no significant pigment variations noted; the color is relatively uniform across the lesion.

Texture and Surface Characteristics: The surface is rough and warty, with a granular texture. There are small, raised bumps clustered together, giving it a candle-wax appearance.

Shape and Border: The lesion has a slightly irregular shape, with a central area that looks more elevated than the surrounding skin. The border is somewhat irregular but not sharply defined.

Additional Relevant Clues: There is no visible inflammation or redness around the lesion. There is no bleeding or scaling noted on the lesion itself.

The lesion exhibits characteristics consistent with Seborrheic Keratosis, including a rough, warty texture, a light brown to grayish-brown color, and an irregular shape. The lack of inflammation or bleeding supports the diagnosis of Seborrheic Keratosis. The appearance of this condition is further illustrated in Figure 2.

Example-3: Clinical features-Based Q&A

User Prompt: What are the clinical features of Seborrheic keratoses?

Seborrheic keratoses can appear on most skin areas, ex- cept palms, soles, and mucous membranes. They vary in size, color, and texture, from flat to raised, smooth to warty. They may be solitary or grouped on areas like the scalp, breasts, spine, or groin, often sticking to the skin.

Ultraviolet (UV) exposure may also contribute to their formation. In some cases, the use of epidermal growth factor receptor (EGFR) inhibitors has been linked to an increased number of these lesions.

*3.1.3. **Preference Conversation Dataset for DPO:***

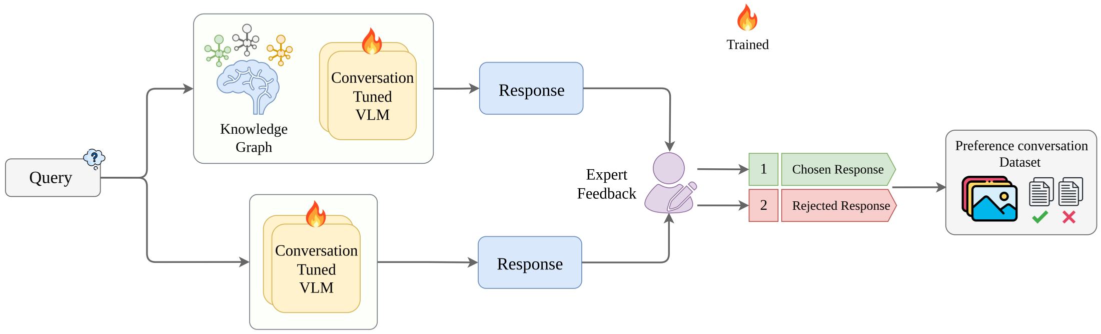

Figure 3: Preference Dataset Creation

In our proposed system, a feedback-based preference dataset is generated to enhance model response quality and adaptability shown in Figure 3. The fine-tuned model generates an initial response, and then a refined version is generated with knowledge graph–based Retrieval Augmented Generation (KG-RAG) assistance. These response pairs are evaluated, often by human annotators, comparing the original and KG-RAG assisted versions. Annotators indicate their preference based on quality, accuracy, and criteria like factual correctness. This collected feedback forms the preference dataset, used for subsequent alignment training. Once both responses are generated, they are used to construct the preference dataset. As expert verification was not available, for our experiments, we designated the KG-RAG–based response as the preferred (chosen) response and the original output as the rejected response. Training the model with this dataset using direct preference optimization (DPO) will help the model internalize the factual responses generated with the aid of KG-RAG. An illustrative example of this dataset is presented in Figure 4.

User Prompt: *"What is the treatment of actinic keratosis?"*

Rejected Response

Actinic keratoses, also known as "liver spots," are rough, scaly patches on sun-exposed skin. They can be treated with various methods depending on their severity and location. Some common treatments include cryotherapy (freezing them), dermabrasion (scraping off dead skin cells using a sharp tool or brush), topical corticosteroids (steroid hormones applied directly to the skin), and systemic antibiotics if the infection has spread beyond the initial spot. It's essential to consult a healthcare professional for personalized advice based on your specific case.

Chosen Response

The main treatment for actinic keratosis is excision, which involves surgically removing the affected skin lesion. This can be done using various methods, including:

- **Surgical excision:** Removes the entire lesion and some surrounding tissue to prevent spread.
- **Mohs surgery:** A layered removal technique that preserves healthy skin while ensuring full excision.
- **Curettage:** Scraping the lesion using a curette; ideal for smaller lesions under local anesthesia.
- **Electrodessication:** Uses an electric current or focused light beam to destroy abnormal cells.

In addition, the following non-surgical methods may be applied:

- **Topical medications:** Such as retinoids or vitamin A to slow abnormal cell growth.
- **Photodynamic therapy:** Activates a light-sensitive drug to kill abnormal cells.
- **Laser therapy:** A non-invasive method that uses concentrated light to eliminate affected tissue.

Figure 4: Comparison between a rejected and a KG-RAG–chosen response for an example question

3.2. *Training Stages*

3.2.1. *Stage 1: Ground Truth Prompt-Guided Reinforcement Learning with AGRPO*

To emulate expert dermatologist reasoning, which involves visual assessment, and medical knowledge, in a vision language model, we developed **Adaptive-GRPO (AGRPO)**. Our method trains the model to build a logical argument for its diagnosis, similar to a clinician. In Grouped Relative Policy Optimization (GRPO), a Base VLM generates multiple responses. A reward function scores these responses, granting higher rewards for correct diseases. Advantages, calculated based on these scores, update the Base VLM to improve future reasoning. A reference model, often the initial base model, stabilizes training using KL Divergence.

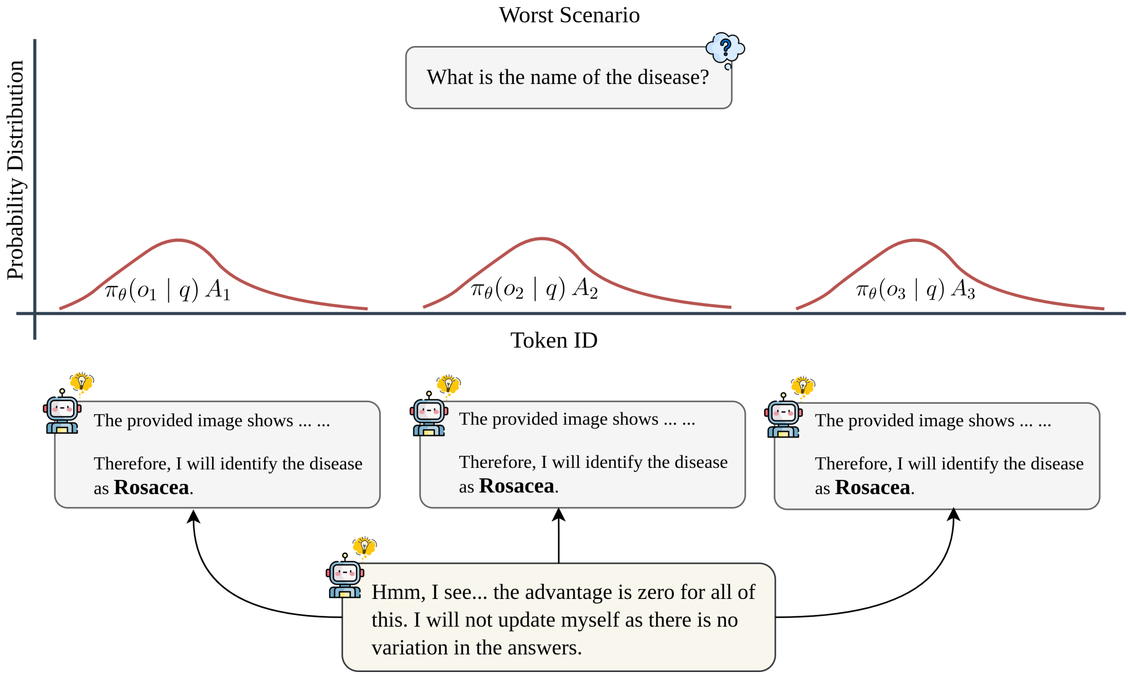
Worst Scenario
What is the name of the disease?
Probability Distribution
$\pi_\theta(o_1 \mid q)\, A_1$
$\pi_\theta(o_2 \mid q)\, A_2$
$\pi_\theta(o_3 \mid q)\, A_3$
Token ID
The provided image shows ... ... Therefore, I will identify the disease as **Rosacea**.
The provided image shows ... ... Therefore, I will identify the disease as **Rosacea**.
The provided image shows ... ... Therefore, I will identify the disease as **Rosacea**.
Hmm, I see... the advantage is zero for all of this. I will not update myself as there is no variation in the answers.

Figure 5: Worst case scenario for GRPO in low generation setting

GRPO struggles significantly in low-resource settings (e.g., low GPU RAM or a small number of response generations per query), which leads to poor performance. GRPO typically requires generating a large number of candidate outputs (32 to 64) per input query to ensure that at least one correct answer is present and can provide a useful advantage signal [24]. However, in low-resource setups, the number of generations is often limited to only 2 to 5 due to GPU memory constraints. With so few generations, the model frequently fails to detect the correct disease and often predicts the same disease across all outputs. As shown in Figure 5, for a given prompt, the model may repeatedly generate the same incorrect disease with nearly identical probability distributions across all sampled responses.

To address these limitations and enable effective learning, we propose Adaptive Grouped Relative Policy Optimization (AGRPO). Our method introduces two key strategies: (i) an adaptive ground-truth hint, and (ii) a modified way to calculate the learning signal, called the advantage function. For the first strategy shown in Figure 6, we provide the model with a hint about the disease relevant to the given question, while also instructing it not to rely solely on the hint, but to consider other possible diseases. Importantly, we do not provide hints for every question as this would cause the model to overfit to hints. Instead, hints are given probabilistically during training. This mechanism helps the model generate more diverse answers and, in most cases, at least one correct answer, thereby reducing the likelihood that the advantage signal collapses to zero. For visualization of our method, in Figure 6, we can see that the model provides probability distribution with higher advantage for the correct response for the same prompt along with a hint.

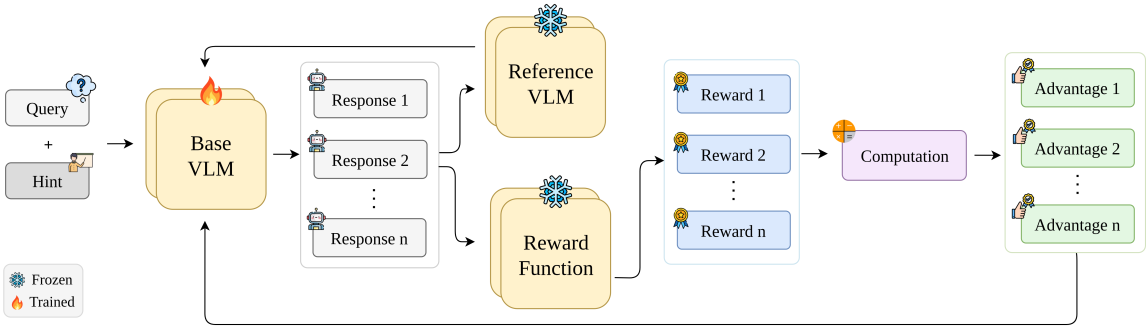

Figure 6: Adaptive Grouped Relative Policy Optimization (AGRPO). Hint inclusion probability will be reduced per epoch at a certain rate.

However, small language models (with fewer than 4 billion parameters) pose an additional challenge. The smaller the model, the less reliably it follows the instructions. For such models, hints sometimes fail to increase diversity, and the model may still generate the same response regardless of whether a hint is provided. Ideally, diverse generations would lead to different rewards, which in turn produce different advantages relative to each other. But when all responses are correct or incorrect, the computed advantage collapses to zero. If KL-divergence regularization is not applied, the loss function also falls to zero, resulting in no gradient update. Without gradient updates, the model receives no useful signal indicating whether the group of responses was collectively good or bad. This limitation highlights a fundamental weakness of GRPO, it performs only intra-group updates (comparing responses within a group), but lacks an update mechanism for the group as a whole when all the responses are same (Does not matter whether they are right or wrong). For example, if all the responses in a group are correct, the model should still receive a positive signal at the group level rather than defaulting to zero. In contrast, if all responses are incorrect, the model should receive a negative group-level signal. Without this adjustment, GRPO learning is still hindered by poor and unstable signals, making it particularly ill-suited for low-resource training conditions and small language models.

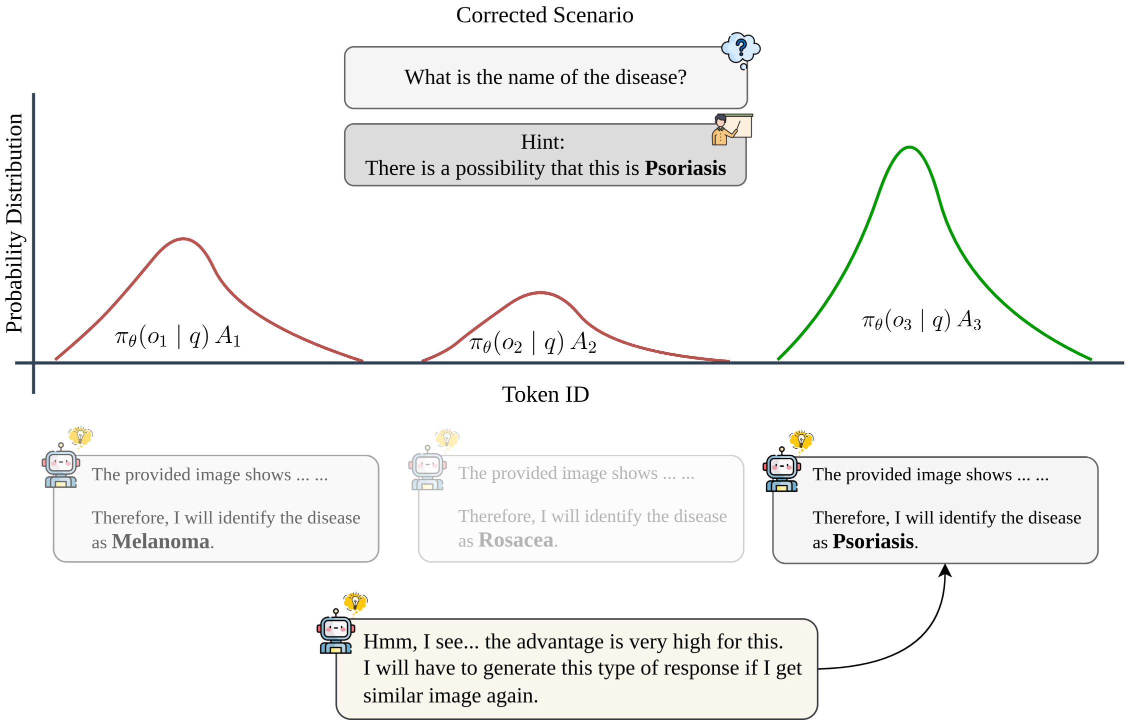
Corrected Scenario
What is the name of the disease?
Hint:
There is a possibility that this is **Psoriasis**
Probability Distribution
$\pi_\theta(o_1 \mid q)\, A_1$
$\pi_\theta(o_2 \mid q)\, A_2$
$\pi_\theta(o_3 \mid q)\, A_3$
Token ID
The provided image shows ... ...
Therefore, I will identify the disease as **Melanoma**.
The provided image shows ... ...
Therefore, I will identify the disease as **Rosacea**.
The provided image shows ... ...
Therefore, I will identify the disease as **Psoriasis**.
Hmm, I see... the advantage is very high for this. I will have to generate this type of response if I get similar image again.

Figure 7: Corrected scenario with AGRPO with low resource setting

The effect of the proposed algorithm is illustrated in Figure 7. By incorporating a hint into the same image–question pair, the model is able to generate a broader range of responses. This, in turn, increases the relative advantage of the correct response and encourages the remaining responses to converge toward it. Importantly, the method also demonstrates effectiveness in scenarios where GRPO fails by uniformly suppressing the entire group.

To address the limitations of standard GRPO, we introduce an adaptive update mechanism that integrates hint-augmented and original contexts in a probabilistic manner. The key idea is to allow the policy to benefit from hints during the early stages of training while gradually reducing its dependence on them as learning progresses. This is achieved by mixing two update strategies: one conditioned on the original prompt and the other on the hint-augmented prompt. The mixture is controlled by a decaying Bernoulli variable, which enables the policy to transition smoothly from hint-guided learning to autonomous reasoning. The objective function is defined as a reformulation of the vanilla GRPO algorithm introduced in [37]:

$$
\begin{aligned}
&\mathcal{J}_{\text{AGRPO}}(\pi_\theta) \\
&= \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{m}\sim\pi_{\theta_{\text{old}}}(\cdot|q),Z\sim\text{Bernoulli}(p_k)} \\
&\Bigg[Z\cdot\frac{1}{m}\sum_{i=1}^{m}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left\{\frac{\pi_\theta(o_{i,t}|q+g_t,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q+g_t,o_{i,<t})}\hat{A}_{i,t},\text{clip}\left(\frac{\pi_\theta(o_{i,t}|q+g_t,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q+g_t,o_{i,<t})},1-\epsilon,1+\epsilon\right)\hat{A}_{i,t}\right\} \\
&\quad+(1-Z)\cdot\frac{1}{m}\sum_{i=1}^{m}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left\{\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t},\text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})},1-\epsilon,1+\epsilon\right)\hat{A}_{i,t}\right\}\Bigg],
\end{aligned}
\tag{1}
$$

where k is the iteration number and $Z \sim \text{Bernoulli}(p_k)$ with $p_k = \gamma^k$ $(p_0 \in (0,1], 0 < \gamma < 1)$ determines the decaying probability of using the hint in the context. As the objective shows, if $Z = 1$ (with probability p_k), the policy update is based on the probability of generating tokens given the **hint-augmented** prompt $(q+g_t)$. If $Z = 0$ (with probability $1-p_k$), the update uses the original prompt (q). Both terms employ a PPO-style clipped importance sampling ratio applied to the advantage $\hat{A}_{i,t}$. The probability p_k decays exponentially with iteration k, ensuring the model relies less on the hint over time, which is key to its adaptive nature. Our modified advantage function combines signals from the rewards to provide a better update, as used in both terms of the objective function $\mathcal{J}_{\text{AGRPO}}$. For a generated output i with reward r_i, the advantage $\hat{A}_i$ is calculated using both its reward relative to others in the generated batch response and absolute reward value:

$$\hat{A}_i = \alpha\frac{r_i-\bar{r}}{\sigma_r+\epsilon}+\gamma r_i$$

Here, $\bar{r}$ and σ_r are the mean and standard deviation of rewards across all generated outputs in the batch, α and γ are weights determining how much each part contributes, and ϵ is a small number added for calculation stability. Algorithm 1 outlines the training process using A-GRPO. In each iteration (outer loop l), the algorithm samples a batch of queries, generates multiple outputs for each query using the current policy $(\pi_{\theta_{old}})$, and calculates a reward for each output using the reward model (r_ϕ). Crucially, it then computes the group relative advantage $(\hat{A}_{i,t})$ based on these rewards. The policy model (π_θ) is updated multiple times (inner loop t) by maximizing the A-GRPO objective (Equation 1). This optimization step involves sampling the Bernoulli variable Z to decide whether to train using the original or hint-augmented prompt context, thus incorporating the adaptive hinting strategy. The reward model (r_ϕ) is also continuously trained using a replay mechanism, improving its ability to score responses over time.

We apply this A-GRPO technique to the base Vision-Language Model named Qwen2.5-VL. Upon completion of this stage, we obtain a model that is specialized in visual disease detection, which is named as Reasoning Aligned VLM in Figure 1, capable of interpreting visual cues from medical images and reasoning towards an accurate diagnosis. The model's reasoning is guided by two key reward functions:

1. Format Adherence Reward: Encourages the model to reason logically and follow a structured approach before making a diagnosis. A reward of $+1$ is given for proper format adherence, while a penalty of -1 is applied for failure to follow the prescribed structure.

2. Detected Disease Reward: The "Detected Disease Reward" critically guides towards accurate diagnoses while strongly penalizing harmful misclassifications. It uses a multi-faceted

Algorithm 1 Adaptive Group Relative Policy Optimization

```
**Input:** initial policy model $\pi_{\theta_{\text{init}}}$, reward models $r_\phi$, task prompts $\mathcal{D}$, hyperparameters $\epsilon, \beta, \mu$
**for** iteration $l = 1, \ldots, I_{\text{do}}$ **do**
    reference model $\pi_{\text{ref}} \leftarrow \pi_\theta$
    **for** step $s = 1, \ldots, M$ **do**
        Sample a batch $\mathcal{D}_b$ from $\mathcal{D}$
        Update the old policy model $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
        Sample $G$ outputs $\{o_i\}_{i=1}^{m} \sim \pi_{\theta_{\text{old}}}(\cdot|q)$ for each question $q \in \mathcal{D}_b$
        Compute rewards $\{r_i\}_{i=1}^{m}$ for each sampled output $o_i$ by running $r_\phi$
        Compute $\hat{A}_{i,t}$ for the $t$-th token of $o_i$ through group relative advantage estimation
        **for** GRPO iteration $t = 1, \ldots, T$ **do**
            Update the policy model $\pi_\theta$ by maximizing the A-GRPO objective (Equation 1)
            Update $\pi_\phi$ through continuous training using a replay mechanism
        **end for**
    **end for**
**end for**
**Output:** $\pi_\theta$
```

approach combining correctness and error severity. Predefined constants handle general outcomes (e.g., correct prediction, failure), while a crucial severity matrix assigns penalties for specific true/incorrect disease pairings. This structure prioritizes diagnostic accuracy and patient safety by accounting for misclassification risks. Full quantitative details are in Table 2.

3.2.2. Stage 2: Dermatological Concept & Knowledge Alignment

This first phase generates a disease detection specialized model known as the Reasoning Aligned VLM as shown in Figure 1. The Reasoning Aligned VLM is skilled at understanding visual hints from medical images and reasoning to an accurate diagnosis. But a dermatologist's involvement would go beyond simple visual identification. Having recognized a skin condition from its appearance, they explain what it is, discuss potential causes, and prescribe treatment. While the Reasoning Aligned VLM is at its strongest in the visual reasoning phase, it lacks the broader clinical knowledge required to conduct discussions on causes or treatments, as these aspects go beyond direct visual data analysis and require access to stored information. To address this shortcoming and equip the Reasoning Aligned VLM with conversational ability, we performed **Stage 2: Supervised Fine-Tuning (SFT)** on **Reasoning Aligned VLM** using an **Image Conversation Dataset**, thereby arriving at **Conversation Tuned VLM**. The primary objective of this SFT step was to allow the resulting conversation Tuned VLM to represent internalized clinical knowledge.

3.2.3. Improving Diagnostic Accuracy with Knowledge Graph Integration and Preference-Based Tuning

Dermatologists reference medical texts, research databases, and clinical guidelines to diagnose complex skin conditions and ensure treatments align with current knowledge, including rare cases. To reduce inaccurate or hallucinated responses—a critical issue in dermatology—we employ a Knowledge Graph-based Retrieval-Augmented Generation (KG-RAG) approach. This system uses a structured knowledge graph of entities, relationships, and facts (e.g., symptoms, causes, treatments) to supply verified medical knowledge. When a user submits a query, the model extracts clinical concepts, retrieves relevant graph triples, and generates fact-based responses.

To emulate how dermatologists internalize experience for quicker decisions, we enhance the finetuned model from Stage 2 using **Direct Preference Optimization (DPO)**. While KG-RAG improves accuracy, it can increase computational load by expanding the context window. DPO helps the model prioritize knowledge-based responses without relying on KG-RAG for familiar concepts. As shown in Figure 3, a preference dataset is created using two versions of the model—one with KG-RAG and one without. For each image-query pair, both models generate outputs. The KG-RAG response, favored for accuracy, is labeled **chosen**; the other is **rejected**. This dataset trains the model with DPO loss, enabling it to prefer knowledge-based reasoning and internalize medical patterns. The resulting model, **DermIQ-VLM**, mirrors a dermatologist's expertise—delivering reliable answers efficiently and with lower computational cost (see Figure 1).

4. Experiments and Results

In the absence of established benchmark datasets for the dermatological domain, we crafted a custom dataset of 35 image-text pairs sourced from authentic dermatological resources. This dataset serves as our benchmark for evaluating the performance of the system in dermatological disease detection.

4.1. ***Experimental Setup***

For **Stage-1: AGRPO**, training was performed on a multi-GPU system with two T4 GPUs (15GB each). Memory and computational savings were achieved using 4-bit quantization and LoRA, which limited trainable parameters. The LoRA configuration set rank=32 and lora_alpha=64, targeting the `q_proj`, `k_proj`, `v_proj`, and `o_proj` modules with 0.05 dropout. Training parameters included a learning rate of 1e-5, a batch size of 1 per device with 4 gradient accumulation steps, 2000 training steps, 10 epochs, and a temperature of 0.9. Generation parameters were set to 3 for the 2B model and 2 for the 3B model due to GPU constraints.

For **Stage-2: Fine-Tune** and **Stage-3: DPO** (as shown in Figure 1), 4-bit quantization was again utilized for memory efficiency. LoRA was applied with rank=8, lora_alpha=16, and 0.05 dropout. LoRA target modules were expanded to include `down_proj`, `o_proj`, `k_proj`, `q_proj`, `gate_proj`, `up_proj`, and `v_proj`. Training used a learning rate of 1e-5 over 2 epochs, with a batch size of 1 per device and 2 gradient accumulation steps. The AdamW optimizer (weight decay=0.01), a linear scheduler (warmup ratio=0.03), and gradient clipping (max norm=0.3) were employed.

4.2. ***Baseline Methods and Evaluation Metric***

We compared our DermIQ-VLM with the baseline models such as Qwen2.5-VL-2B-Instruct, Qwen2.5-VL-3B-Instruct, LLaVA, LLaVAMed, etc. For disease detection, the text between `<answer>` and `</answer>` tags is extracted and matched with the ground truth answer. Majority voting involves taking multiple predictions from a model and selecting the most frequent answer as the final output to enhance prediction reliability.

For analysis of conversation quality, we employed the most advanced large language models - Grok and GPT-4 - as judges. These models were prompted to check if responses were generated in the given format, particularly whether the text adhered to the boundary tags `<thinking>`... `</thinking>` and `<answer>`...`</answer>`. The quality of generated responses is thereby evaluated using various metrics, including **factual accuracy**, which evaluates the accuracy of model output for identifying the disease; **relevance**, measuring the degree to which the answer attaches to the user's query; and **completeness**, determining whether the answer covers all aspects that are necessary in the diagnosis and reasoning process.

4.3. ***Detected Disease Reward:***

This reward component is pivotal for guiding the model towards accurate diagnostic conclusions and minimizing harmful misclassifications. Its calculation is based on a comprehensive set of predefined constants for various prediction outcomes and a detailed severity matrix for specific disease misclassifications. Together, these elements form a nuanced reward system designed to encourage both accuracy and clinically safe predictions. The complete reward structure is detailed in Table 2.

Parameter Description	**Value**
Base reward for correct disease identification	$+10.0$
Bonus for correct identification if no ground truth hint was used	$+3.0$
Penalty for model failing to output a valid disease prediction	-5.0
Reward/Penalty if the ground truth label itself is unknown or invalid	-0.5
Default penalty for misclassifications not covered by severity matrix	-2.5

Table 1: General reward and penalty constants for various prediction outcomes.

True Disease (Abbreviation)	**Incorrectly Predicted Disease**						
	AK	BCC	Derm.	Mel.	Psor.	Ros.	SK
Actinic Keratosis (AK)	N/A	-1.0	-3.0	-1.5	-3.0	-3.0	-2.0
Basal Cell Carcinoma (BCC)	-1.5	N/A	-4.0	-2.0	-4.0	-4.0	-3.0
Dermatitis (Derm.)	-2.5	-3.0	N/A	-3.5	-0.5	-0.7	-2.5
Melanoma (Mel.)	-3.0	-2.5	-5.0	N/A	-5.0	-5.0	-4.0
Psoriasis (Psor.)	-2.5	-3.0	-0.5	-3.5	N/A	-0.8	-2.5
Rosacea (Ros.)	-2.5	-3.0	-0.7	-3.5	-0.8	N/A	-2.5
Seborrheic Keratosis (SK)	-1.0	-2.0	-1.5	-3.0	-1.5	-1.5	N/A

Table 2: Comprehensive Reward Structure for the Disease Detection Task. This table outlines: (A) general reward and penalty constants applied in various prediction scenarios, and (B) the detailed severity-based penalty matrix for specific disease misclassifications.

The reward function (Table 2) structures the learning process by applying specific rewards and penalties based on prediction outcomes. A correct disease identification receives a base reward, and an additional bonus is granted if this diagnosis was made without relying on a ground truth hint (as detailed in Part A). Conversely, if the model predicts an incorrect disease, a penalty is applied. This penalty is critically scaled according to the clinical severity and potential harm of the misdiagnosis as specified in the severity matrix (Part B). For instance, misclassifying Melanoma as Dermatitis incurs a significant penalty of -5.0, whereas misclassifying Actinic Keratosis as Seborrheic Keratosis results in a lesser penalty of -2.0, thus strongly encouraging the model to avoid high-risk misclassifications. Furthermore, various other prediction outcomes are handled, including a penalty for no valid prediction, a reward for unknown ground truth cases, and a default mismatch penalty, all listed in Part A.

4.4. Performance of Disease Detection with Reasoning

To evaluate the impact of system-level prompting and reinforcement learning (RL) tuning, we compared the performance of existing zero-shot vision-language models (VLMs) with our proposed **DermIQ-VLM** across different configurations. Figure 8 presents user prompt and the model's step-by-step reasoning process.

Type	**Backbone Model**	**Samples Trained**	**Weighted F1 Score (%)**	**Weighted Precision (%)**	**Weighted Recall (%)**
Existing Zero-shot VLMs	Qwen2-VL-2B	–	20.37	21.98	21.62
	Qwen2.5-VL-3B	–	55.42	58.26	54.05
	LLaVA-1.5-7b	–	23.81	23.81	27.14
GRPO-tuned VLM	Qwen2-VL-2B	800	32.36	53.26	35.14
	Qwen2.5-VL-3B		60.98	64.41	59.46
DermIQ-VLM	Qwen2.5-VL-3B	800	**62.48**	**66.54**	**62.16**

Table 3: Single Shot Evaluation Performance (Summary) with merged and two-line Model Type cells (simple box style).

This table 3 shows the performance of all evaluated models with a single prediction. Our DermIQ-VLM-3b model, trained on 800 samples using our AGRPO method, achieved a Weighted F1-score of 62.48% and Weighted precision of 66.54%. This is the best performance among all models evaluated in this single-shot setting. Specifically, **DermIQ-VLM** surpassed the strongest existing zero-shot model, Qwen 2.5 3B (F1: 55.42%, precision: 58.26%). Furthermore, it outperformed the same Qwen 2.5 3B base model when tuned with a standard GRPO approach (F1: 60.98%, precision: 64.41%). This demonstrates that our AGRPO tuning method leads to better results, even with limited training data.

Type	**Backbone Model**	**Samples Trained**	**Weighted F1 Score (%)**	**Weighted Precision (%)**	**Weighted Recall (%)**
Existing Zero-shot VLMs	Qwen2-VL-2B	–	28.80	33.52	35.14
	Qwen2.5-VL-3B		53.36	61.08	54.05
	LLaVA-1.5-7b		32.54	31.90	43.24
GRPO-tuned VLM	Qwen2-VL-2B	800	53.65	54.10	56.76
	Qwen2.5-VL-3B		60.48	64.95	62.16
DermIQ-VLM	Qwen2.5-VL-3B	800	**62.28**	**68.27**	**64.86**

Table 4: Majority Voting Evaluation Performance (Summary) with merged and two-line Model Type cells (simple box style).

This table 4 shows the overall performance when majority voting is used across several predictions. Our DermIQ-VLM-3b model, trained on 800 samples, is the best performer with a Weighted F1-score of 62.28% and Weighted precision of 68.27%. This confirms its reliability. Majority voting generally improved performance for most models compared to their single-prediction results, showing that combining multiple predictions helps. DermIQ-VLM-3b also outperformed the GRPO-tuned Qwen2.5 3B model (F1: 60.48%, precision: 64.95%), further highlighting the benefits of our AGRPO tuning approach.

A crucial finding from our evaluation is the direct comparison between our proposed DermIQ-VLM-3b (tuned with AGRPO) and the same base model (Qwen 2.5 3B) tuned with a standard GRPO approach. In the single-shot inference setting (Table 3), DermIQ-VLM-3b demonstrates a significant uplift in performance compared to the GRPO-tuned Qwen 2.5 3B. It achieves a Weighted F1-score of 62.48% compared to GRPO-tuned Qwen 2.5 3B's 60.98%, representing an approximate 2.5% improvement in F1-score. Similarly, its Weighted precision of 66.54% surpasses GRPO's 64.41%, showing an approximate 3.3% gain. The superior performance extends to the majority voting scenario as well (Table 4). Here, DermIQ-VLM-3b attains a Weighted F1-score of 62.28% against GRPO's 60.48%, marking an approximate 3.0% increase. For Weighted precision, DermIQ-VLM-3b achieves 68.27% compared to GRPO's 64.95%, an approximate 5.1% improvement. This consistent outperformance across both evaluation settings underscores the effectiveness and enhanced diagnostic capability achieved by our advanced Adaptive GRPO (AGRPO) reinforcement learning tuning strategy over a conventional GRPO approach, even when starting from an identical base VLM.

Detailed Classification Reports Per Disease

This detailed table provides individual performance metrics for each model when making a single prediction. For **GRPO-tuned Qwen2.5 3B**, it showed good F1-scores for Melanoma (0.89), Basal Cell Carcinoma (0.67), and Rosacea (0.67), performing similarly to the best un-tuned model, Qwen 2.5 3B, in these areas. Our proposed **DermIQ-VLM** (AGRPO-tuned) performed better than all other models for Dermatitis (F1: 0.73) and Rosacea (F1: 0.80). It also had perfect precision for Melanoma (1.00) and perfect recall for Basal Cell Carcinoma (1.00). Overall, **DermIQ-VLM** showed strong F1-scores for Basal Cell Carcinoma (0.73), Dermatitis (0.73), Melanoma (0.75), and Rosacea (0.80), demonstrating its ability to accurately identify specific conditions quickly. However, the un-tuned Qwen 2.5 3B still achieved the highest F1-scores for Actinic Keratosis (0.67), Basal Cell Carcinoma (0.80), and Psoriasis (0.62) in this single-prediction setting.

Disease	**Existing Zero-shot VLMs**									**GRPO-tuned VLM**			**DermIQ-VLM**		
	Qwen2-VL-2B			**Qwen2.5-VL-3B**			**LLaVA-1.5-7B**			**Qwen2.5-VL-3B**			**Qwen2.5-VL-3B**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
AK	0.20	0.20	0.20	0.75	0.60	0.67	0.00	0.00	0.00	0.50	0.60	0.55	0.60	0.60	0.60
BCC	0.33	0.75	0.46	0.67	1.00	0.80	0.00	0.00	0.00	0.60	0.75	0.67	0.57	1.00	0.73
DER	0.20	0.20	0.20	0.33	0.20	0.25	0.33	0.20	0.25	0.33	0.20	0.25	0.67	0.80	0.73
MEL	0.00	0.00	0.00	1.00	0.80	0.89	0.33	0.20	0.25	1.00	0.80	0.89	1.00	0.60	0.75
PSO	0.60	0.38	0.46	0.62	0.62	0.62	0.50	0.50	0.50	0.57	0.50	0.53	0.67	0.50	0.57
ROS	0.00	0.00	0.00	0.75	0.60	0.67	0.50	1.00	0.67	0.75	0.60	0.67	0.80	0.80	0.80
SK	0.00	0.00	0.00	0.40	0.40	0.40	0.00	0.00	0.00	0.33	0.40	0.36	0.33	0.20	0.25

Table 5: Single Shot Evaluation: Precision (P), Recall (R), and F1-Score (F1) per disease for each model.

Furthermore, **DermIQ-VLM** shows competitive F1-scores for Basal Cell Carcinoma (0.73), Dermatitis (0.73), Melanoma (0.75), and Rosacea (0.80). This per-disease performance highlights

Disease	Existing Zero-shot VLMs									GRPO-tuned VLM			DermIQ-VLM		
	Qwen2-VL-2B			Qwen2.5-VL-3B			LLaVA-1.5-7B			Qwen2.5-VL-3B			Qwen2.5-VL-3B		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
AK	0.31	0.80	0.44	0.50	0.60	0.55	0.00	0.00	0.00	0.67	0.40	0.50	0.62	1.00	0.77
BCC	0.20	0.25	0.22	0.40	1.00	0.57	0.00	0.00	0.00	0.67	1.00	0.80	0.67	1.00	0.80
DER	0.29	0.40	0.33	0.50	0.20	0.29	0.50	0.20	0.29	0.40	0.40	0.40	1.00	0.40	0.57
MEL	0.00	0.00	0.00	1.00	0.80	0.89	0.33	1.00	0.50	0.83	1.00	0.91	1.00	1.00	1.00
PSO	0.45	0.62	0.53	0.62	0.62	0.62	0.71	0.62	0.67	0.55	0.75	0.63	0.45	0.62	0.53
ROS	1.00	0.20	0.33	1.00	0.40	0.57	0.38	1.00	0.56	1.00	0.40	0.57	0.67	0.40	0.50
SK	0.00	0.00	0.00	0.20	0.20	0.20	0.00	0.00	0.00	0.50	0.40	0.44	0.50	0.20	0.29

Table 6: Majority Voting Evaluation: Precision (P), Recall (R), and F1-Score (F1) per disease for each model.

DermIQ-VLM's ability to accurately diagnose specific conditions even with a single inference. Notably, with majority voting, the **GRPO-tuned VLM (Qwen2.5 3b)** performed strongly. It had the best F1-scores for Psoriasis (0.63), Rosacea (0.57), and Seborrheic Keratosis (0.44). It also matched **DermIQ-VLM**'s perfect recall and an F1 of 0.80 for Basal Cell Carcinoma. However, our **DermIQ-VLM** (RL-tuned) achieved a perfect F1-score for Melanoma (1.00) and the best F1-scores for Actinic Keratosis (0.77) and Dermatitis (0.57). Its perfect recall for Actinic Keratosis (1.00), Basal Cell Carcinoma (1.00), and Melanoma (1.00), along with high precision for Dermatitis (1.00) and Melanoma (1.00), means it can identify many true cases and avoid false cases, which is crucial in a medical context.

? **User Prompt:** *"Describe the image and find out the disease."*

✓ **Model Response**

<thinking>
Dermatological Analysis:
Color and Pigmentation: The image shows reddened areas on the face, which could be indicative of Rosacea.
Texture and Surface Characteristics: The reddened areas appear smooth and without visible scales or ulcers, which is consistent with Rosacea.
Shape and Border: The reddened areas have irregular borders, which is characteristic of Rosacea.
</thinking>
<answer>
Rosacea
</answer>

Figure 8: An example of the reasoning quality of our model

Figure 9 compares the training performance of two qwen models, each trained with GRPO and AGRPO. It shows that, with same configurations and same number of response generation(e.g. 2 responses per query), the AGRPO algorithm (represented by the red and orange lines) consistently outperforms GRPO (represented by the green and blue lines), with smoother and higher rewards as training progresses. The models with larger parameters, like Qwen2.5-3B-VL achieve better performance compared to the smaller Qwen2-2B-VL models. Overall, the GRPO algorithm leads to better training results, with a clear improvement in mean combined reward as the training steps increase, especially after 800 steps.

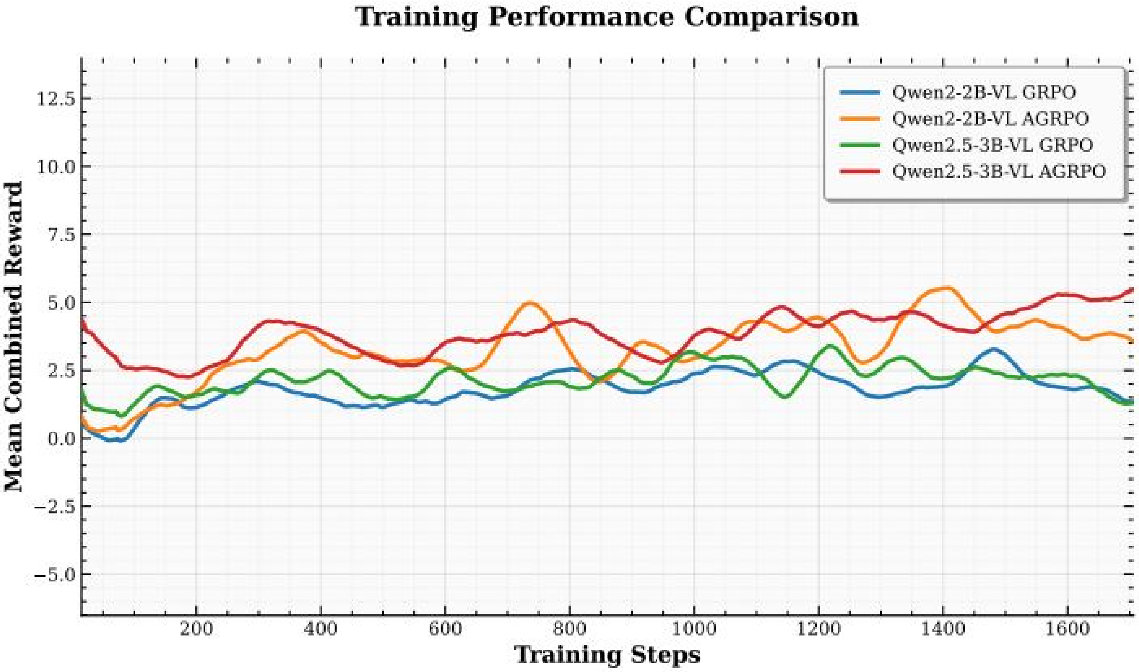

Figure 9: Reward comaprison of two different models trained on GRPO and AGRPO

4.5. Performance of Conversational Quality

Fine-tuning the vision-language model alone initially produced quite good results, which motivated the integration of Retrieval-Augmented Generation (RAG) supported by a knowledge graph derived from a web-crawled unstructured dataset. We evaluated the performance progression across three models—Base VLM(Qwen 3B), Conversation-Tuned VLM and DermIQ-VLM using human evaluations by Grok and GPT-4-Turbo, focusing on factual accuracy, relevance, and completeness (Table **??**).

Conversation-Tuned VLM shows a better result over base vlm with consistently higher average scores across all metrics. When assessed by Grok, scores for Conversation-Tuned VLM were 7.75 (accuracy), 7.33 (relevance), and 7.10 (completeness), while those for Base VLM were recorded at 6.05, 6.65, and 6.10, respectively. GPT-4-Turbo also showed the same trend in preferring Conversation-Tuned VLM with scores of 6.95 (accuracy), 6.60 (relevance), and 6.43 (completeness), compared to Base VLM's 5.13, 5.53, and 5.10. Clearly, these results show that fine-tuning on a conversational dataset considerably improves response quality.

With further refining through Knowledge Graph-based Direct Preference Optimization (DPO), we observed even greater advances. The final model, **DermIQ-VLM**, scored the highest across all evaluation sets—Grok rated it 8.53 (accuracy), 8.42 (relevance), and 8.45 (completeness), while

Backbone Model	Type	Topic	Accuracy			Relevance			Completeness		
			Grok	**GPT-4**	**Diff.**	**Grok**	**GPT-4**	**Diff.**	**Grok**	**GPT-4**	**Diff.**
Qwen2.5-VL-3B	Pretrained	Treatment	6.1	5.3	0.8	6.2	5.1	1.1	5.2	4.4	0.8
		Causes	4.2	3.7	0.5	6.1	5.0	1.1	5.1	4.2	0.9
		Demographics	6.0	5.2	0.8	6.2	5.3	0.9	6.1	5.2	0.9
		Features	7.9	6.3	1.6	8.1	6.7	1.4	8.0	6.6	1.4
		Average	**6.05**	**5.13**	**0.92**	**6.65**	**5.53**	**1.12**	**6.10**	**5.10**	**1.00**
	Conversation Tuned	Treatment	7.9	7.2	0.7	8.0	7.1	0.9	8.0	7.1	0.9
		Causes	8.1	7.3	0.8	8.1	7.4	0.7	7.1	6.5	0.6
		Demographics	7.0	6.2	0.8	5.1	4.7	0.4	6.2	5.7	0.5
		Features	8.0	7.1	0.9	8.1	7.2	0.9	7.1	6.4	0.7
		Average	**7.75**	**6.95**	**0.80**	**7.33**	**6.60**	**0.73**	**7.10**	**6.43**	**0.67**
	DermIQ-VLM	Treatment	8.6	7.7	0.9	8.1	7.4	0.7	8.6	7.8	0.8
		Causes	8.4	7.6	0.8	8.5	7.7	0.8	8.3	7.6	0.7
		Demographics	8.2	7.4	0.8	8.3	7.5	0.8	8.2	7.5	0.7
		Features	8.9	7.8	1.1	8.8	7.9	0.9	8.7	7.8	0.9
		Average	**8.53**	**7.63**	**0.90**	**8.42**	**7.63**	**0.79**	**8.45**	**7.68**	**0.77**

Table 7: Evaluation of conversational responses using Accuracy, Relevance, and Completeness metrics, judged by Grok and GPT-4, with differences (Diff.) highlighting scoring bias.

GPT-4-Turbo ratings improved to 7.63, 7.63, and 7.68, respectively. Importantly, the gap between Grok and GPT-4 evaluations remained fairly small, indicating that the improvements were robust across the judgment criteria. Completeness showed a similarly stable upward trend, with the scoring bias narrowing to 0.77 in the final model. This figure 10 demonstrates the conversation quality of our model:

? **User Prompt:** *"What is the treatment of Rosacea?"*

✓ **Model Response**

<thinking>
This image shows rosacea. Treatment options for rosacea include:
- **Retinoids:** These can help reduce inflammation and improve skin texture.
-**Corticosteroids:**Topical corticosteroids can provide short-term relief of redness and swelling.
- **Topical antibiotics** These can help control bacterial overgrowth in the skin.
- **UVA therapy:** This can be used alone or in combination with other treatments.
- **PUVA therapy:** A combination of UVA light and a photosensitizer like psoralen.
- **Antibiotics:** Oral antibiotics can be used if rosacea is severe.
- **Anti-inflammatory drugs:** These can be used for severe cases.
</thinking>
<answer>
Rosacea is treated with therapies, antibiotics, corticosteroids, phototherapy, systemic medications, surgical procedures, and lifestyle modifications.
</answer>

Figure 10: An example of the response quality of our model

5. Conclusion

This study addresses the challenges of developing medical vision-language models (VLMs) tailored for low-resource settings, focusing on the detection and discussion of skin diseases with explainable, factually accurate, and structured reasoning for dermatological applications. We propose DermIQ-VLM, a vision-language model designed to emulate the diagnostic process of dermatologists. To achieve this, we introduce a memory-efficient variant of Group Relative Policy Optimization (GRPO) alongside multi-step training strategies, including GRPO, supervised fine-tuning, Knowledge Graph Retrieval Augmented Generation (KG-RAG), and Direct Preference Optimization (DPO). These methods significantly enhance diagnostic performance, achieving an average detection rate of 4% with very limited training data. The generated responses were further evaluated by benchmark large language models for factual accuracy, relevance, and completeness. While these results are promising, further improvements could be realized by curating larger, higher-quality datasets and expanding the knowledge base. Future work will focus on broadening observations and refining preference learning strategies. In summary, this research marks a significant step toward developing interpretable and clinically impactful AI tools to support dermatologists in resource-constrained real-world settings.

6. Acknowledgement

We would like to acknowledge that we have utilized AI models to assist in refining our grammatical mistakes and enhancing the conversational flow throughout our research. These AI tools have played a valuable role in improving the clarity and coherence of our work, ensuring a more polished presentation of ideas and communication.

References

[1] Z. Lin, D. Zhang, Q. Tao, D. Shi, G. Haffari, Q. Wu, M. He, and Z. Ge, *Medical Visual Question Answering: A Survey*, *Artificial Intelligence in Medicine*, vol. 143, p. 102611, 2021.

[2] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler, *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions*, *Scientific Data*, vol. 5, no. 1, pp. 1–9, 2018.

[3] A. Hurst, A. Lerer, A.P. Goucher, A. Perelman, A. Ramesh, A. Clark, A.J. Ostrow, A. Welihinda, A. Hayes, A. Radford, and A. Mądry, *GPT-4o System Card*, *arXiv preprint arXiv:2410.21276*, 2024.

[4] xAI, *Grok 3: The Smartest AI in the World*, *xAI News*, Available at: `https://x.ai/news/grok-3`, Accessed: July 9, 2025.

[5] P. Shojaee, I. Mirzadeh, K. Alizadeh, M. Horton, S. Bengio, and M. Farajtabar, *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, *arXiv preprint arXiv:2506.06941*, 2025.

[6] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun, *Dermatologist-level classification of skin cancer with deep neural networks*, *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[7] S. Ghosh, C. K. R. Evuru, S. Kumar, D. Aneja, Z. Jin, R. Duraiswami, and D. Manocha, *A Closer Look at the Limitations of Instruction Tuning*, *arXiv preprint arXiv:2402.05119*, 2024.

[8] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, *Large Language Models in Medicine*, *Nature Medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.

[9] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, and D. Zhou, *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824–24837, 2022.

[10] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Wang, Jialin Ge, Wenbin Liu, and others, *Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution*, *arXiv preprint arXiv:2409.12191*, 2024.

[11] Zhiyong Chen, Fuyong Liu, Chantal Rosenbaum, Peijie Leo, Drew F. K. Williamson, Tiffany Y. Chen, and others, *A visual-language foundation model for computational pathology*, *Nature Medicine*, vol. 29, no. 10, pp. 2653–2664, 2023.

[12] Alon Jacovi and Yoav Goldberg, *Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?*, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4198–4205, 2020.

[13] Michael Moor, Kuan Huang, and others, *Foundation models for generalist medical artificial intelligence*, *Nature*, vol. 621, no. 7980, pp. 37–45, 2023.

[14] D. Sanmartin, *Kg-RAG: Bridging the Gap Between Knowledge and Creativity*, *arXiv preprint arXiv:2405.12035*, 2024.

[15] Y. Kim, H. Jeong, S. Chen, S. S. Li, M. Lu, K. Alhamoud, J. Mun, *et al.*, *Medical Hallucinations in Foundation Models and Their Impact on Healthcare*, *arXiv preprint arXiv:2503.05777*, 2025.

[16] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert, *Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning*, *arXiv preprint arXiv:2502.19634*, 2025.

[17] Titus J. Brinker, Achim Hekler, Alexander H. Enk, Joachim Klode, Axel Hauschild, Carola Berking, and others, *Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic image classification task*, *European Journal of Cancer*, vol. 113, pp. 47–54, 2019.

[18] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, and others, *Scaling Instruction-Finetuned Language Models*, *arXiv preprint arXiv:2210.11416*, 2022.

[19] Shenglin Zhang, Pengtian Zhu, Minghua Ma, Jiagang Wang, Yongqian Sun, Dongwen Li, Jingyu Wang, Qianying Guo, Xiaolei Hua, Lin Zhu, and others, *Enhanced Fine-Tuning of Lightweight Domain-Specific Q&A Model Based on Large Language Models*, *2024 IEEE 35th International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pp. 61–66, 2024.

[20] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen, *Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference*, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3428–3448, 2019.

[21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, and others, *Chain-of-thought prompting elicits reasoning in large language models*, *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 24824–24837, 2022.

[22] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, and others, *Towards Generalist Biomedical AI*, *arXiv preprint arXiv:2307.14334*, 2023.

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee, *Visual Instruction Tuning*, *arXiv preprint arXiv:2304.08485*, 2023.

[24] X. Zhang, H. Sun, Y. Zhang, K. Feng, C. Yang, and H. Meng, *Critique-GRPO: Advancing LLM Reasoning with Natural Language and Numerical Feedback*, *arXiv preprint arXiv:2506.03106*, 2025.

[25] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, and others, *Retrieval-augmented generation for knowledge-intensive NLP tasks*, *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.

[26] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec, *QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering*, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 535–546, 2021.

[27] Daniel N. Nicholson and Casey S. Greene, *Constructing knowledge graphs and their biomedical applications*, *Computational and Structural Biotechnology Journal*, vol. 18, pp. 2857–2866, 2020.

[28] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, and others, *Survey of hallucination in natural language generation*, *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[29] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn, *Direct preference optimization: Your language model is secretly a reward model*, *Advances in Neural Information Processing Systems*, vol. 36, pp. 53728–53741, 2023.

[30] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn, *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*, *arXiv preprint arXiv:2305.18290*, 2023.

[31] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, and others, *Training language models to follow instructions with human feedback*, *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 27730–27744, 2022.

[32] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and others, *Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning*, *arXiv preprint arXiv:2501.12948*, 2025.

[33] Yuhao Zhang, Hang Xu, Jiacheng Liu, Xiaohan Lu, Qingfang Liu, and Dong Xu, *Large-scale Radiology Report Generation via General Multi-modal Pretraining*, *arXiv preprint arXiv:2212.13547*, 2022.

[34] Aaron Kline, Haolin Wang, Yucheng Li, Shane Dennis, Meghan Hutch, Zhen Xu, Fei Wang, Feixiong Cheng, Yuan Luo, *Multimodal machine learning in precision health: A scoping review*, *npj Digital Medicine*, vol. 5, no. 1, pp. 171, 2022.

[35] Shuyi Wang, Zihao Zhao, Xi Ouyang, Tong Liu, Qian Wang, Dinggang Shen, *Interactive computer-aided diagnosis on medical image using large language models*, *Communications Engineering*, vol. 3, no. 1, pp. 133, 2024.

[36] Fei Gan, Lei Chen, Wei Qin, Qian Han, Xiang Long, Hao Fan, Xiang Li, Hong Yu, Jinxia Zhang, Na Xu, Jie Cheng, *Ophthus-gpt: Multimodal ai for automated reporting in ophthalmic b-scan ultrasound*, *medRxiv*, pp. 2025-03, 2025.

[37] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y.K., Wu, Y. and Guo, D., 2024. *Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300*