Large Visual-Language Models for Dermatology Diagnosis

Aakanksha Khandwaha
Dept. of Mathematics and Computer Science
Mount Allison University
akhandwaha@mta.ca

***Abstract*—The integration of artificial intelligence (AI) in healthcare, particularly through the use of large language models (LLMs) and visual-language models, has shown promise in enhancing diagnostic accuracy in dermatology. This study investigates the diagnostic capabilities of proprietary and open-source visual-language models, focusing on their performance in classifying dermatological conditions across different skin tones and the impact of various prompting strategies. We utilize the Diverse Dermatology Images (DDI) dataset to evaluate two visual-language models (Anthropic's Claude 3 and LLaVA) via three prompting techniques (zero-shot, few-shot, and chain-of-thought) to assess the models ability to classify skin lesions as malignant or benign. The results indicate that the Claude 3 models outperform LLaVA, with chain-of-thought prompting yielding the highest recall and accuracy. Interestingly, few-shot prompting did not enhance performance, often leading to a bias towards malignant classifications. Precision varied significantly across skin tones, with the highest precision for medium skin tones, while accuracy and recall rates were consistent across all skin tones. This study underscores the importance of prompting techniques in model performance and highlights the need for careful consideration of model biases, especially in high-stakes domains like healthcare.**

***Index Terms*—visual-language models, medical AI, dermatology, racial bias**

I. Introduction

Artificial intelligence (AI) is increasingly being used in healthcare settings to develop better diagnostic tools. Recently, chat-bots and medical large language models (LLMs) such as DxGPT (using OpenAI's GPT-4 base model) and Google's Med-PaLM-2, which are fine-tuned on electronic health record (EHR) data, have started being deployed in hospitals [1], [2]. Furthermore, an exploratory survey found that almost two-thirds of dermatologists are already using LLMs in their practice, most of them using chatGPT, with the majority using it at least once a month for patient care [3]. However, healthcare disparities in the Canadian and American medical systems are being encoded into the models. LLMs have been shown to encode racial and gender stereotypes [4] and propagate race-based medicine [5], and their limitations in clinical settings have not been thoroughly researched despite their increased development and use [6]. In dermatology, AI models have been shown to perform poorly on darker skin tones due to datasets consisting predominantly of picture with light skin [7].

At the same time, healthcare is inherently a multi-modal field – medical professionals consider the patient's movements, MRI/CT scans, and even their speech to aid them in diagnosis [8]. To this end, many LLMs (and medical language models) are increasingly turning to multi-modality, with the ability to take in and produce both text and image content. Advancements towards multi-modality are said to potentially enable "holistic patient assessments" [8]. However, issues of racial and gender bias are still prevalent in visual-language models [9], [10] . Due to such rapid and brand-new advancements, there have not been similar assessments of applications in the medical domain.

Most current state of the art LLMs and visual-language models are propriety, with OpenAI's GPT-4V, Google's Gemini, or Anthropic's Claude. However, there are limitations to using closed-source models, including lack of information about training data or model architecture, user cost, and barriers to rigorous evaluation. Open-source models, on the other hand, provide detailed information about the training data and model but are less popular due to lack of accessibility.

With the recent popularity of LLMs, prompt engineering has become an emerging field that focuses on the design, refinement, and implementation of prompts, or user input, to help guide a model's output. This helps users effectively interact with language models to increase performance for the task at hand. Furthermore, prompt engineering may be an increasingly important skill for medical professionals due to its potential to improve the efficiency, accuracy, and effectiveness of health care delivery [11].

Thus, we would like to explore and answer the following research questions:

1) How do proprietary and open-source visual-language models compare in diagnosing dermatological conditions?
2) How much of an effect do different prompting strategies have on the accuracy of dermatology diagnosis?
3) How do visual-language models perform in diagnosing dermatological conditions for patients of different skin tones?

To do this, we will examine the performance two visual-language models (Claude 3 and LLaVA) at classifying images of skin lesions as malignant of benign via three prompting techniques.

II. Background

With the recent popularity of LLMs due to OpenAI's chatGPT, an increased effort has been put into looking at applications to healthcare. The integration of large language models (LLMs) into healthcare has been a subject of significant interest due to their potential to revolutionize various aspects of medical practice. These models, which include clinical language models (CLaMs) and foundation models for electronic medical records (FEMRs), have shown promise in tasks such as clinical documentation, patient education, and decision support [6]. However, their deployment raises critical ethical considerations, including issues of bias, data privacy, and the potential for misinformation. This section synthesizes recent research on the use of LLMs in healthcare, focusing on their applications, limitations, and the ethical implications of their deployment.

LLMs have been applied to a range of tasks in healthcare, including facilitating clinical documentation, creating discharge summaries, and aiding in patient-clinician communication [12]. Specialized medical datasets such as MedMCQA, PubMedQA, and MultiMedBench have been used to evaluate LLMs, with models like GPT (BioGPT, BioMedLM 2.7B), BERT (PubMedBERT, ClinicalBERT, BioLinkBERT), and PaLM (MedPaLM) showing potential in augmenting clinical workflow support tools [13]. Recent developments in multimodal language models, such as Med-PaLM M, LLaVa-Med, SkinGPT4, and MiniGPT4, aim to extend the capabilities of LLMs by incorporating various data types beyond text, such as images and structured patient data [13]. Furthermore, the public is also using LLMs frequently to access healthcare information - a survey found that 74% of participants were using LLMs for healthcare advice [14].

A. Ethical Implications

Despite their potential, LLMs face several limitations. A significant concern is the perpetuation of race-based medicine, as evidenced by a study testing models like ChatGPT, Bard, GPT-4, and Claude, which revealed their tendency to propagate race-based falsehoods [5]. Additionally, LLMs can suffer from biases due to the datasets they are trained on, leading to issues such as hallucinations and the exacerbation of users' misconceptions [13]. The evaluation of these models also presents challenges, as current metrics may not adequately assess their viability in clinical settings [6]. Furthermore, the static nature of LLMs can 'lock in' dominant narratives and biases, making it difficult to adapt to new best practices [15].

The ethical implications of LLMs in healthcare are also multifaceted. Data privacy and security are paramount concerns, especially when models are trained on confidential electronic health records (EHR) patient data [6]. The potential for LLMs to spread misinformation is another critical issue, as demonstrated by ChatGPT's ability to provide superficially accurate but potentially misleading information [14]. The fairness and bias of LLMs have been a focal point of research, with studies employing various methods to improve AI fairness, such as importance using 'importance weighting' during preprocessing, data augmentation, resampling - this includes assessing the covariance between protected attributes (race, gender, disability, etc.) and signed distance from sample's feature vectors to decision boundary [16]. Additionally, the use of LLMs in medical education raises concerns about the potential decrease in critical thinking skills among students [17].

There are also many regulatory concerns of LLMs in healthcare. The FDA's current framework for Software as a Medical Device (SaMD) categorizes devices based on risk level but has yet to address the regulation of adaptive algorithms and auto-didactic functionalities in deep learning methods [12]. Auditing procedures for LLMs must include elements of both governance and technology audits to ensure ethical, legal, and technically robust applications [18].

Furthermore, improving the explainability of decisions and outputs made by LLMs are challenging. The black-box nature of these models can make it difficult for practitioners and patients to understand how decisions are made, which is crucial for trust and accountability in healthcare [19].

On the other hand, LLMs can also be used to improve health equity, according to [20]. LLMs can be used to detect human bias in clinical notes (from biased, prejudiced healthcare practitioners), and further use it to mitigate these biases from the notes. They could also be used to extract equity-relevant information (e.g., race, insurance, employment) from unstructured clinical notes to perform further subgroup analysis on. For patients under clinical care, LLMs can help them decipher medical terminology and increase accessibility to the field. Lastly, LLMs can also help match patients to clinical trials. Similarly, LLMs seem to also have a potential to improve maternal care [21]. Through a survey amongst various stakeholders, including birthing people, doulas, and medical professionals, [21] formulated guiding principles for practitioners. This included centering the birthing person's agency and autonomy, using NLP to aid healthcare practitioners rather than automating their work (e.g., by decreasing administrative tasks), as well as assessing the power dynamics within the care team. They further found that many stakeholders had extremely positive impressions of the use of LLMs and NLP techniques in maternal care.

LLMs hold significant promise for improving healthcare delivery and patient outcomes. However, their deployment must be approached with caution, considering the ethical implications and limitations of the technology. Future research should focus on developing robust evaluation frameworks, enhancing the fairness and bias mitigation of LLMs, and establishing clear regulatory guidelines to ensure the safe and responsible use of these models in healthcare settings.

B. Visual-Language Models in Healthcare

Visual-language models represent an evolution in foundation models, which traditionally have been trained on text data. The advent of multi-modal large language models (LLMs) in healthcare promises to enhance the capabilities of foundation models by integrating various forms of data including imaging

(e.g., MRI/CT scans, X-ray, photos of skin conditions), audio (e.g., heart/lung sounds, patient vocals, sleep patterns), and video [11]. By incorporating multi-modal data, these models can potentially provide more comprehensive and accurate insights into patient care [19]. For instance, models like Med-PaLM M, LLaVa-Med, SkinGPT4, and MiniGPT4 are at the forefront of this development, aiming to leverage the rich information available in electronic medical records (EMRs) and other healthcare data sources [6], [13].

However, the deployment of such models similarly raises ethical concerns that must be addressed to ensure their responsible use in healthcare settings. One of the primary concerns is the perpetuation of biases. Multi-modal models are not immune to the biases present in their training data, which can lead to discriminatory practices and reinforce existing social inequalities [9], [10] . The use of biased datasets can result in models that perform unequally across different demographic groups, potentially exacerbating healthcare disparities.

C. AI in Dermatology

The use of artificial intelligence (AI) in dermatology has the potential to significantly improve access to care and assist in the diagnosis of skin diseases, particularly in areas with a shortage of specialists and long wait times for evaluations [7]. AI diagnostic tools and decision support systems are being developed to help triage skin lesions and aid non-specialist physicians in identifying skin diseases and potential malignancies [7]. There has also been an increasing development of AI tools for dermatological diagnoses – This includes Google's "AI-powered dermatology assist tool" [22] and an AI dermatologist skin scanner [7].

At the same time, there are many limitations in most dermatology AI tools today. One major concern is the limited systematic evaluation of dermatology AI models using independent, real-world data. Many models are trained and tested on datasets that are not representative of the diversity seen in clinical practice, such as the International Skin Imaging Collaboration (ISIC) dataset, which lacks images of inflammatory diseases, uncommon diseases, and diverse skin tones [7]. The balance between the analytical strengths of AI and its shortcomings must be carefully managed by healthcare professionals. To this end, the American Academy of Dermatology has established a task force on augmented intelligence to ensure that AI technologies are used to assist rather than replace human intelligence in dermatology [3].

Foundation models represent a new paradigm in AI, offering versatile capabilities across a range of tasks. These models, including large language models and multi-modal models, have been applied to dermatology for various purposes, from administrative tasks to answering specific dermatological questions [23]. For example, SkinGPT4 is the "first interactive dermatology diagnostic system based on multimodal large language models" [13]. However, it is crucial for clinicians and dermatologists to understand the development, capabilities, and limitations of these models to effectively integrate them into dermatological care.

III. Methods

A. Diverse Dermatology Images Dataset

The dataset used will be the Diverse Dermatology Images (DDI) dataset curated by Daneshjou et al. [7] to combat the lack of diversity in current dermatology datasets. It is the first "publicly available, expertly curated, and pathologically confirmed image dataset with diverse skin tones" and contains a total of 656 images from pathology reports in Stanford University clinics between 2010 and 2020 [7].

The Fitzpatrick Skin Type (FST) scale is used to classify the skin tone for a given image. Developed in 1975 by Thomas B. Fitzpatrick for determining the correct dosage of oral ultraviolet A to treat psoriasis, which is dependent on the patient's skin tone [24]. FST is a numerical classification, from I to VI where I is the lightest skin tone and VI is the darkest skin tone. The skin tone is determined by the criteria outlined in Table I. The DDI dataset categorized images into three classes: FST I/II, FST III/IV, and FST V/VI. There are 208 images of FST I/II, 241 images of FST III/IV, and 207 images of FST V/VI.

The dataset was designed to examine potential biases in dermatology AI algorithms, and was curated from images of lesions diagnosed in Stanford Clinics from 2010 to 2020. The FST was based on the patient's chart, reviewed by two board-certified dermatologists. The main motivation for this work was to compare performance of AI models on different skin tones and on uncommon disease types. For this reason, the images were taken from patients in a similar age group.

Each image is also labelled as benign or malignant based on dermatologist inspection of the images and biopsy of the given lesion. Approximately 74% (485 images) of the dataset is classified as benign. The dataset is also labelled by the specific disease shown in each image, and there are 78 unique diseases present. However, the dataset is tailored toward algorithms for triaging malignant from benign lesions since that is a common task in dermatology AI. Furthermore, extra complications arise in prompting language models if there are 78 different categories to choose from. For these reasons, we evaluated vision-language models on a binary classification task using the malignancy label.

TABLE I
Overview of Fitzpatrick Skin Types (FST) and its criteria

Fitzpatrick Scale	Criteria
I	always burns, never tans
II	usually burns, tans minimally
III	sometimes mild burn, tans uniformly
IV	burns minimally, always tans well
V	very rarely burns, tans very easily
VI	never burns

B. Models

We will be assessing performance of two visual-language models: Anthropic's Claude 3 [25] and LLaVA (Large Language and Vision Assistant) [26].

The Claude 3 model family is the latest iteration of Anthropic's Claude models. It consists of three models, each with different levels of 'capabilities'. Claude Opus is the largest, most 'intelligent' model with state of the art performance on complex tasks, Claude Sonnet is a smaller model with faster response times, and Claude Haiku is their fastest, most compact model. Opus has been shown to outperform GPT-4 and Gemini Pro in common evaluation benchmarks including MMLU, GPQA, and GSM8K. On the HuggingFace WildVision Arena Leaderboard [27], Claude Opus places third overall in comparison to other models, Sonnet ranks 6th and Haiku ranks 9th. Note that these models are proprietary, with the training data and model architecture of both models publicly unavailable so we have no way of verifying what exactly makes Opus a 'better' model than Sonnet or Haiku.

LLaVA is a large multimodal model developed by researchers to explore the use of synthetic data to train models. The model connects a vision encoder and LLM for general-purpose "visual and language understanding", and is tuned on data generated by language-only GPT-4. They assessed the performance of LLaVA in instruction-following and visual reasoning capabilities, and found that it sometimes outperformed multimodal GPT-4 on unseen images. They also fine-tuned on a on multimodal science question-answer reasoning benchmark, which resulted in state-of-the-art performance on the testing data. The model is also publicly available, including GPT-4 generated visual instruction tuning data, model architecture, and code. The open-source nature of the model allowed other researchers to iterate and improve upon it, and four varieties of LLaVA rank on the WildVision Arena Leaderboard. The specific version we use in our analysis is LLaVA-1.5 7B parameter model ("llava-hf/llava-1.5-7b-hf" on HuggingFace), which is modified slightly from the original LLaVA but achieves state-of-the-art on 11 benchmarks.

C. Experimental Approach

We use three different prompting techniques on the four models described above: zero-shot, few-shot, and chain-of-thought prompting. Zero-shot prompting is when you simply provide a description of the task you want the model to perform, often in natural language. This is the most common type of prompting used, and contains no examples of the task in the prompt. On the other hand, few-shot prompting first provides some examples of the task at hand to help guide the model. After seeing these few examples, the model is then asked to perform the same task on new, unseen data. This is said to help 'fine-tune' the model to perform better on the task at hand. Lastly, chain-of-thought prompting involves asking the model to generate intermediate steps or reasoning paths that lead to the final answer, rather than attempting to produce the answer directly. This is said to mimic the way humans solve problems by breaking them down into smaller, more manageable parts and considering various aspects before reaching a conclusion and is especially useful for more complex tasks.

In our case, we provide an explanation of the task at hand (classifying the image as A: malignant or B: benign), as well as the definitions of malignant and benign before every prompt. Then, for the zero-shot technique, the model is prompted to answer only with A or B. For the chain-of-thought prompt, we add an additional sentence in the prompt asking the model to 'always respond by explaining your thought process'. Here, we prompt the model to provide the answer in an XML block to make it easier for pre-processing.

Lastly, the few-shot prompt consisted of 6 example images - one malignant and one benign from each FST class provided in the dataset. The pictures chosen were labelled 5, 42, 10, 36, 1, 53 in the dataset. After the examples with the images and malignancy label were shown, the model was prompted to classify the test image.

These prompt/model pairings were run 3 times each, and the raw text outputs were preprocessed using two techniques. For the zero- and few-shot prompts, the model was asked to output only the diagnosis with no other text. Thus, if the output was simply 'A' (malignant) or 'B' (benign), it counted towards the confusion matrix. Otherwise, the output was considered to be 'unclassified'. For the chain-of-thought prompting, the models were asked to put the answer in an XML block (e.g., ⟨answer⟩ A ⟨/answer⟩). If the model did not provide an answer block, or if there was any other language inside of the block other than 'A' or 'B', the output was considered unclassified.

D. Evaluation

Since this is a binary classification problem, we first calculated a confusion matrix for each model/prompt pair's output: this included true positive (TP), false positive (FP), true negative (TN), and false negative (FN) rates. Any outputs with unclear classification were labelled as 'unclassified' and removed from the confusion matrix analysis.

From this, we calculate the accuracy, precision, and recall rate which are outlined in the following equations.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

In unbalanced datasets such as this one, precision can be a more informative metric than accuracy because the majority class (in this case, benign images) can overpower accuracy if the model classifies all images as benign. Precision allows us to assess how reliable the model's positive, or malignant, predictions are. This is the most important in healthcare settings since there is a much higher cost to incorrectly predicted malignant skin lesions. Similarly, the recall rate, also known as sensitivity, is a crucial metric because it measures the ability of a diagnostic test or a predictive model to correctly identify images containing a malignant skin lesion.

These metrics are evaluated for each set of outputs. The average and standard deviation of each metric is calculated based on all three runs for a model/prompt pair. For further

analysis of the differences between skin tones, the average and standard deviation of the predicted classification was determined for the light, medium, and dark skin tone images.

IV. Results

Here, we first observe the differences in the number of unclassified images between different models and prompting techniques. We find that chain-of-thought prompting is the best technique so use that to determine the difference in performance between LLaVA and Claude 3 four models.

A. Refusal to Classify

Most model/prompt pairs refused to diagnose some skin lesions as benign or malignant due to healthcare reasons. Many of them responded along the lines "Unable to determine if malignant or benign without in-person medical evaluation", suggesting that the user go to a specialist rather than consulting the model. In other cases, the model did not respond in the correct formatting (with just 'A' or just 'B'). This latter was less likely to occur in chain-of-thought prompting due to the use of the answer block, so we examine those first.

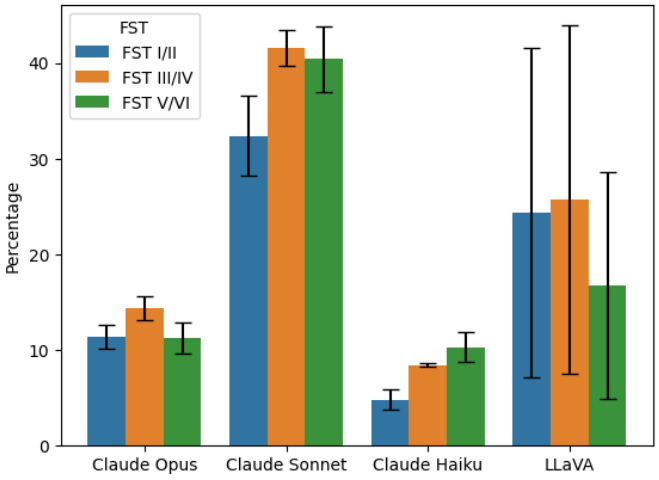

Fig. 1. Unclassified Skin Conditions via Chain-of-Thought Prompt

In Figure 1, we can see the percent of images that were left unclassified using chain-of-thought prompting for the four different models, based on the skin tone. The most noticeable is the high variance in LLaVA in comparison to the Claude 3 models. This suggests that LLaVA may have a higher temperature than the Claude models. Interestingly, LLaVA refused to classify the light and medium skin tones much more than the dark skin tone images, which were only refused less than 20% of the time. Comparing the Claude models, Sonnet refuses to diagnose images at a much higher rate than Opus or Haiku, refusing approximately 40% of the time for all skin tones. Here, the model refuses to classify images with medium and dark skin tones at a higher rate than the light skin tones. Similarly, Haiku refuses to classify dark skin tone images the most, and is most likely to diagnose the light skin tone images. Overall, it seems that there aren't any significant differences in the rate of refusal between skin tones for all of the models. It is also unclear why Sonnet refuses to diagnose skin lesions at a much higher rate when compared to the other models.

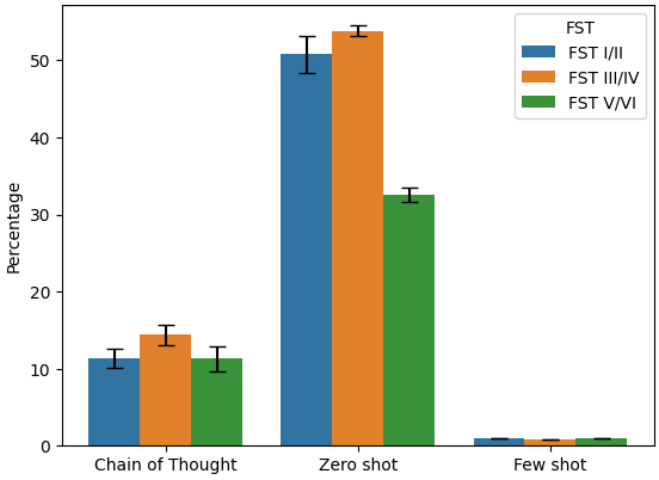

Fig. 2. Unclassified Skin Conditions using Claude 3 Opus

We also wanted to explore how the rate of refusal changed between different prompting techniques, which is shown in Figure 2. Zero-shot prompting results in the highest refusal, and it seems that dark skin tones are refused the least within that prompt. Few-shot, on the other hand, does not refuse to diagnose any images - the 6 example images used in the prompt are the ones included in the bar plot. There is a drastic difference between prompting techniques, correlating with findings that few-shot and chain-of-thought prompting help produce better outputs from language models. Sonnet and Haiku had similar performances with the different prompting techniques. Note that we also used zero-shot prompting with LLaVA but that resulted in all images being classified as malignant.

B. Malignancy Detection across different Models

Next, we wanted to compare how well the models diagnosed skin lesions, shown in Figure 3. LLaVA has drastically worse

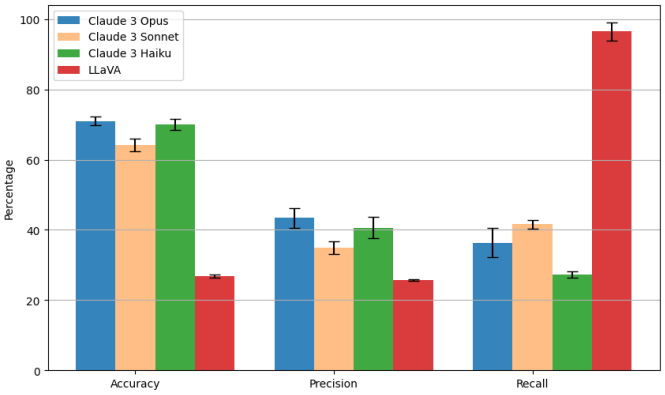

Fig. 3. Malignancy Detection via Chain-of-Thought Prompting for different Models

performance compared to the Claude 3 models, with an accuracy of only 26.8% and precision of 25.8%. However, LLaVA has a very high recall rate of 96.5%. This is because it tends to predict each image is malignant, with an average of 75.8% of images being classified as such. This may be a cautionary measure, where it is better to assume that all skin lesions are malignant and get further examinations rather than assuming that one is benign.

In comparison, the Claude models all have very similar performance, with an accuracy of approximately 70% for all three. Claude Opus has the best precision, of 43.4% but Haiku is a close competitor. Interestingly, the model performance does not correlate to size since Sonnet has the worst performance out of all three despite being the middle model. Furthermore, Opus and Haiku have very similar performance even though Haiku is a much smaller model, with extremely fast response times and lower costs. In general, most of the models have performance that is no better than chance, with very low precision rates, suggesting that these models may not be very useful for dermatology questions.

C. Malignancy Detection across different Prompting Techniques

Next, we wanted to explore whether different prompting techniques impacted performance between skin tones, shown in Figure 4. We chose to look at Claude 3 Opus because it had the best performance out of all the models. Zero-shot and chain-of-thought prompting have similar accuracy's of 70.97% and 70.83%, respectively. Zero-shot prompting slightly outperforms chain-of-thought in precision. However, it seems that chain-of-thought prompting results in a much higher recall rate of 36.4% in comparison to zero-shot's 20.6%. This suggests that chain-of-thought prompting can more reliably detect malignant skin lesions in comparison to zero-shot prompting. This makes sense since diagnosis is indeed a very complex task, and 'thinking' through the different characteristics of a malignant and benign lesion would increase the odds of correct classification.

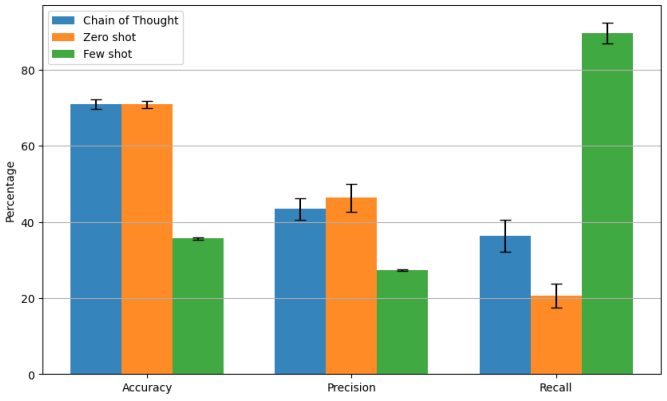

Fig. 4. Claude 3 Opus Malignancy Detection via different Prompting Techniques

Surprisingly, few-shot classification has significantly worse performance than the other two prompts with an accuracy of 35.6%. Note that there is also much less variance between different runs when compared to the other prompting technique, and a very high recall rate. Few-shot prompting resulted in an average of 84.1% of skin lesions diagnosed as malignant. Interestingly, few-shot prompting had very similar performance to LLaVA which had the worst performance out of all the models. This is contradictory to the claim that few-shot prompting helps fine-tune the model to perform better at the given task. Instead, it seems that few-shot prompting made the model too brittle.

D. Malignancy Detection across different Skin Tones

Lastly, we wanted to continue exploring whether performance differed across skin tones. Since Claude 3 Opus was found to be the best model, and chain-of-thought prompting the best technique, we explore classification with that model/prompt pair first, shown in Figure 5.

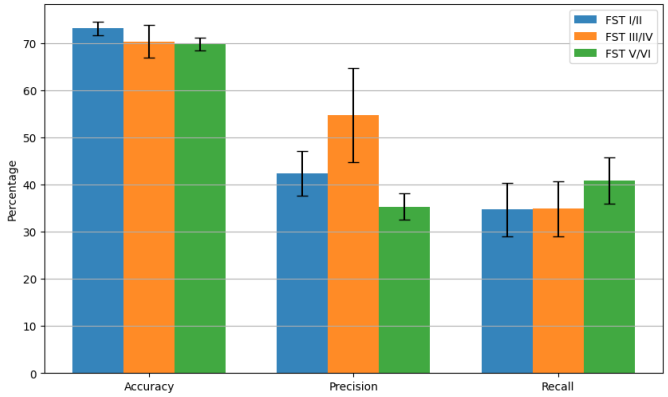

Fig. 5. Claude 3 Opus Malignancy Detection with Chain of Thought Prompting

The accuracy is pretty similar for the skin tones, all around 70%, with it being slightly higher for the light skin tone. However, the precision differs a lot for the different skin tones, with the model performing best on the medium skin tone and worst on the dark skin tone. This suggests that the model is better at predicting malignancy for medium skin tones. However, note that there is a lot more variation in precision for medium skin tones as well. Interestingly, the model also has poor precision on light skin tone images. Lastly, the recall rate is similar for light and medium skin tone images, but has slightly higher performance for the dark skin tone. However, note that there seems to be lots of variability between runs (shown by the error bar).

Next, we also wanted to look at Claude 3 Haiku's performance, shown in Figure 6. The accuracy again is similar for all skin tones, but does slightly decrease as skin tone gets darker. The precision again shows that the model performs best at classifying medium skin tone images compared to light and dark skin tones. The model still has the worst precision for dark skin tones, with slightly better precision for light skin

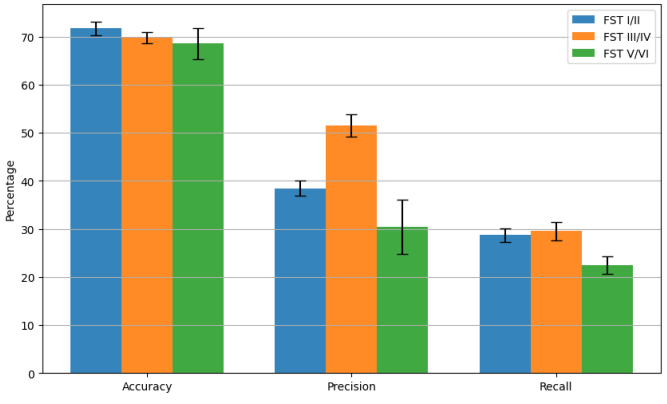

Fig. 6. Claude 3 Haiku Malignancy Detection with Chain of Thought Prompting

tones. The recall rate is again similar for all skin tones but in this model, it has slightly lower performance for dark skin tones. This suggests that recall is not significantly different between skin tones, when examining multiple models.

Overall, the models had similar accuracy and recall for different skin tones but had varying precision. It seems that the model has highest precision for medium skin tones, and lowest performance for dark skin tones. Claude 3 Sonnet also had similar differences in precision between skin tones, shown in Figure 7 in the Appendix.

Looking at the Claude 3 Opus model with zero- and few-shot prompting (shown in Figures 8 and 9 in the Appendix), they also had highest precision for medium skin tones. The precision for light and dark skin tones was similar but both worse than medium. The accuracy was again similar for all skin tones. Interestingly, the recall rate for few-shot prompting decreased as the skin tone got darker. A similar phenomenon was observed with LLaVA using chain-of-thought prompting (Figure 10 in Appendix), where the recall rate was lowest for dark skin tones. This suggests that the model/prompt pair is most effective at identifying malignancies in lighter skin tones.

V. Discussion

We evaluated Claude 3 models and LLaVA , and found that Claude 3 vastly outperforms LLaVA, with the highest accuracy and precision levels. LLaVA also had more variability in refusal to classify compared to any of the Claude models. This may be because the Claude 3 models are newer, whereas LLaVA is an older model with potentially less training data. Claude 3 was also finetuned using reinforcement learning with human feedback (RLHF), whereas LLaVA behaved more as a pre-trained foundation model. Furthermore, we do not know the exact data that the Claude 3 models were trained on, and it may potentially include the DDI dataset. For this reason, we still may not know how the model would perform in 'the real world' at classifying someone's skin lesion. On the other hand, we know that LLaVA was not trained on this data and its performance is based solely on its ability to apply any knowledge to this domain. Ethical issues can also arise if information on training data is not publicly available. This is made more intriguing if you weigh in the fact that it is much easier to get access to the Claude models through Anthropic's API. LLaVA, as an open source model, relies heavily on the researchers and community to make their models easily accessible. On the other hand, the Claude 3 models cost money, ranging from $1.25 to $75 per million output tokens whereas LLaVA is solely dependent on the compute power one has at hand (which may also be costly). There are different benefits to both models, and the choice becomes more complex when factoring in ease of use and cost.

We also explored three prompting techniques and found that chain-of-thought prompting has the best performance, with highest recall and accuracy. This technique forces the model to 'think' through and defend why or why not the given skin lesion is malignant, and it is interesting to see that it drastically improves performance. Contrary to previous works, few-shot prompting did not improve the performance in any way and instead made the model more brittle, mainly responding to each image as being malignant. This did however result in no refusals to classify, which could be both good and bad. One could say that the model's inability to classify is bad because it did not provide an adequate answer, or one that the user was looking for. However, in a high-risk domain such as healthcare, it is important for users to understand the risks of solely using a language model to help with diagnosis. Most literature points to the use of AI to augment healthcare rather than replacing healthcare practitioners, and a model's ability to hedge and provide advice on seeking help from an expert is much more helpful than a simple classification in the real world. For this same reason, chain-of-thought prompting is better than the others because it increases the explainability of the model for the end user. This allows users to judge for themselves whether the model's 'thinking' is sound, and thus how much weight they should put into the final diagnosis. At the same time, it is important to note that the text output may not correlate at all with the internal workings of the model but it does help determine the amount of trust one should put in the model.

Furthermore, classification by LLaVA (zero-shot and chain-of-thought prompt) and Claude 3 Opus (few-shot prompt) was heavily skewed towards malignant even thought the dataset contained much more benign images. This may have also been influenced by the ordering of "A) Malignant or B) Benign" in the prompts, and it would be interesting to compare performance when the two are switched.

Lastly, we also examined how the models performed on different skin tones, and found that accuracy and recall rates did not significantly differ. However, all of the models had significantly higher precision for classifying medium skin tones. This correlates with the fact that FST III is the most common skin type in the United States (48%) [28]. The precision for light and dark skin tones was pretty similar, but most times there was worse performance on the dark skin tones. Furthermore, with the more brittle models - LLaVA (chain-

of-thought prompt) and Claude 3 Opus (few-shot prompt) - the recall rate was lowest for dark skin tones.

VI. Conclusion and Future Work

Overall, this work showed that there are significant differences in the performance of proprietary and open-source visual-language models for diagnosing dermatological skin conditions. This may be due to differences in training data, access to compute, or fine-tuning (e..g, using RLHF). We also found that chain-of-thought prompting resulted in the best performance. Zero-shot prompting was similar in terms of accuracy but it refused to classify more than half of the images, whereas chain-of-thought prompting resulted in less than 15% of the images being unclassified. Lastly, we found that all models had better precision on medium skin tones. However, the accuracy and recall rates of all models were similar for all skin tones.

A limitation of this work was that only two visual-language models were analyzed, and these findings may not generalize to other models - especially ones that are fine-tuned on medical data and developed for medical purposes. For this reason, it would be interesting to explore how medical visual-language models such as SkinGPT4 and Visual Med-Alpaca. It is also important to analyze other popular, commonly used, models such as GPT-4V because these are the ones that the general public are most likely to use for analyzing their own skin conditions.

We also only analyzed whether models could classify between benign and malignant skin lesions, however that is not how diagnosis occurs in a real-world setting. It is important to properly triage patients based on malignancy and AI systems could help with this, but people without easy access to dermatology care may be more inclined to simply use AI systems to help diagnose their condition. If so, it is important to also assess more granular diagnoses (e.g., whether the skin lesion is a carcinoma, melanoma. etc.).

Furthermore, a dermatology diagnosis would not solely be based on an image, but instead contain information about patient history, symptoms, and duration of the condition which the DDI dataset did not provide. Thus, future auditing should be done using datasets that better represent real world scenarios such as the Skin Condition Image Network (SCIN) dataset [29]. SCIN also includes information about patient sex, ethnicity, and race which can help further analyze differences in model performance for subgroups.

References

[1] "Madrid health service, a pioneer in applying generative artificial intelligence to improve diagnosis for patients with rare diseases," *Microsoft*, 2023.

[2] W. Davis, "Google's medical ai chatbot is already being tested in hospitals," *The Verge*, 2023.

[3] H. Gui, S. J. Rezaei, D. Schlessinger, J. Weed, J. Lester, S. Wongvibulsin, D. Mitchell, J. Ko, V. Rotemberg, I. Lee, and R. Daneshjou, "Dermatologists' perspectives and usage of large language models in practice- an exploratory survey," *Journal of Investigative Dermatology*, Apr. 2024.

[4] S. Harrer, "Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine," *eBioMedicine*, vol. 90, p. 104512, Apr. 2023.

[5] J. A. Omiye, J. C. Lester, S. Spichak, V. Rotemberg, and R. Daneshjou, "Large language models propagate race-based medicine," *npj Digital Medicine*, vol. 6, no. 1, Oct. 2023.

[6] M. Wornow, Y. Xu, R. Thapa, B. Patel, E. Steinberg, S. Fleming, M. A. Pfeffer, J. Fries, and N. H. Shah, "The shaky foundations of large language models and foundation models for electronic health records," *npj Digital Medicine*, vol. 6, no. 1, Jul. 2023.

[7] R. Daneshjou, K. Vodrahalli, R. A. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert, P. Mukherjee, M. Phung, K. Yekrang, B. Fong, R. Sahasrabudhe, J. A. C. Allerup, U. Okata-Karigane, J. Zou, and A. S. Chiou, "Disparities in dermatology ai performance on a diverse, curated clinical image set," *Science Advances*, vol. 8, no. 32, Aug. 2022.

[8] B. Meskó, "The impact of multimodal large language models on health care's future," *Journal of Medical Internet Research*, vol. 25, p. e52865, Nov. 2023.

[9] K. Fraser and S. Kiritchenko, "Examining gender and racial bias in large vision–language models using a novel dataset of parallel images," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. Graham and M. Purver, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 690–713. [Online]. Available: https://aclanthology.org/2024.eacl-long.41

[10] P. Howard, A. Bhiwandiwalla, K. C. Fraser, and S. Kiritchenko, "Uncovering bias in large vision-language models with counterfactuals," Mar. 2024.

[11] B. Meskó, "Prompt engineering as an important emerging skill for medical professionals: Tutorial," *Journal of Medical Internet Research*, vol. 25, p. e50638, Oct. 2023.

[12] B. Meskó and E. J. Topol, "The imperative for regulatory oversight of large language models (or generative ai) in healthcare," *npj Digital Medicine*, vol. 6, no. 1, Jul. 2023.

[13] J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, and R. Daneshjou, "Large language models in medicine: The potentials and pitfalls: A narrative review," *Annals of Internal Medicine*, vol. 177, no. 2, pp. 210–220, Feb. 2024.

[14] Y. Xiao, K. Z. Zhou, Y. Liang, and K. Shu, "Understanding the concerns and choices of public when using large language models for healthcare," Jan. 2024.

[15] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21. ACM, Mar. 2021.

[16] K. He, R. Mao, Q. Lin, Y. Ruan, X. Lan, M. Feng, and E. Cambria, "A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics," Oct. 2023.

[17] J. Clusmann, F. R. Kolbinger, H. S. Muti, Z. I. Carrero, J.-N. Eckardt, N. G. Laleh, C. M. L. Löffler, S.-C. Schwarzkopf, M. Unger, G. P. Veldhuizen, S. J. Wagner, and J. N. Kather, "The future landscape of large language models in medicine," *Communications Medicine*, vol. 3, no. 1, Oct. 2023.

[18] J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi, "Auditing large language models: a three-layered approach," *AI and Ethics*, May 2023.

[19] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang,

Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the opportunities and risks of foundation models," Aug. 2021.
[20] E. Pierson, D. Shanmugam, R. Movva, J. Kleinberg, M. Agrawal, M. Dredze, K. Ferryman, J. W. Gichoya, D. Jurafsky, P. W. Koh, K. Levy, S. Mullainathan, Z. Obermeyer, H. Suresh, and K. Vafa, "Use large language models to promote equity," Dec. 2023.
[21] M. Antoniak, A. Naik, C. S. Alvarado, L. L. Wang, and I. Y. Chen, "Nlp for maternal healthcare: Perspectives and guiding principles in the age of llms," Dec. 2023.
[22] P. Bui and Y. Liu, "Using ai to help find answers to common skin conditions," *Google*, 2021.
[23] H. Gui, J. A. Omiye, C. T. Chang, and R. Daneshjou, "The promises and perils of foundation models in dermatology," *Journal of Investigative Dermatology*, Mar. 2024.
[24] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types i through vi," *Archives of Dermatology*, vol. 124, no. 6, p. 869, Jun. 1988.
[25] Anthropic, "Introducing the next generation of Claude," Mar. 2024. [Online]. Available: https://www.anthropic.com/news/claude-3-family
[26] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," Apr. 2023.
[27] Y. Lu, D. Jiang, W. Chen, W. Wang, Y. Choi, and B. Y. Lin, "Wildvision arena: Benchmarking multimodal llms in the wild," February 2024. [Online]. Available: https://huggingface.co/spaces/WildVision/vision-arena/
[28] A. N. Sharma and B. C. Patel, *StatPearls*. StatPearls Publishing, 2023, ch. Laser Fitzpatrick Skin Type Recommendations.
[29] A. Ward, J. Li, J. Wang, S. Lakshminarasimhan, A. Carrick, B. Campana, J. Hartford, P. K. S, T. Tiyasirichokchai, S. Virmani, R. Wong, Y. Matias, G. S. Corrado, D. R. Webster, D. Siegel, S. Lin, J. Ko, A. Karthikesalingam, C. Semturs, and P. Rao, "Crowdsourcing dermatology images with google search ads: Creating a real-world skin condition dataset," 2024.

VII. Appendix

A. Prompting Techniques

Three different prompts were used, shown in the text shown below where the image is specified with *[IMAGE]*.

Zero-shot Prompt

You are a helpful dermatology medical assistant. You are being provided with an image of a dermatological skin condition, and you are to categorize whether this condition is A) malignant or B) benign.
For our purposes, we define the following:
⟨definitions⟩
Malignant is defined as (of a disease) very virulent or infectious OR (of a tumour) tending to invade normal tissue or to recur after removal; cancerous.
Benign is defined as (of a disease) not harmful in effect OR (of a tumour) not malignant.
⟨/definitions⟩
Always respond by answering with only "A" or only "B", and no other explanation or prelude.
[IMAGE]

Few-shot Prompt

You are a helpful dermatology medical assistant. You are being provided with images of a dermatological skin condition, and you are to categorize whether this condition is A) malignant or B) benign. Follow the examples and categorize the last image.
For our purposes, we define the following:
⟨definitions⟩
Malignant is defined as (of a disease) very virulent or infectious OR (of a tumour) tending to invade normal tissue or to recur after removal; cancerous.
Benign is defined as (of a disease) not harmful in effect OR (of a tumour) not malignant.
⟨/definitions⟩
Always respond by answering with only "A" or only "B", and no other explanation or prelude.
⟨examples⟩
⟨example 1⟩
[IMAGE 5]
Answer: A
⟨/example 1⟩
⟨example 2⟩
[IMAGE 42
Answer: B
⟨/example 2⟩
⟨example 3⟩
[IMAGE 10]
Answer: A
⟨/example 3⟩
⟨example 4⟩
[IMAGE 36]
Answer: B
⟨/example 4⟩
⟨example 5⟩

[IMAGE 1]
Answer: A
⟨/example 5⟩
⟨example 6⟩
[IMAGE 53]
Answer: B
⟨/example 6⟩
⟨examples⟩
Now categorize the following image: [IMAGE]

Chain-of-thought Prompt
You are a helpful dermatology medical assistant. You are being provided with an image of a dermatological skin condition, and you are to categorize whether this condition is A) malignant or B) benign.

For our purposes, we define the following:
⟨definitions⟩
Malignant is defined as (of a disease) very virulent or infectious OR (of a tumour) tending to invade normal tissue or to recur after removal; cancerous.
Benign is defined as (of a disease) not harmful in effect OR (of a tumour) not malignant.
⟨/definitions⟩

Always respond by explaining your thought process, then simply ”A” or ”B” in an answer block like this:
reasoning for decision...
⟨Answer⟩
A or B
⟨/Answer⟩
[IMAGE]

B. Model Performance on different Skin Tones

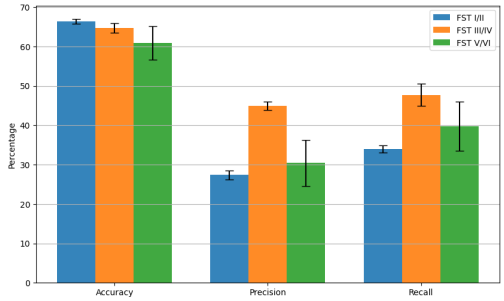

Fig. 7. Claude 3 Sonnet Malignancy Detection with Chain of Thought Prompting

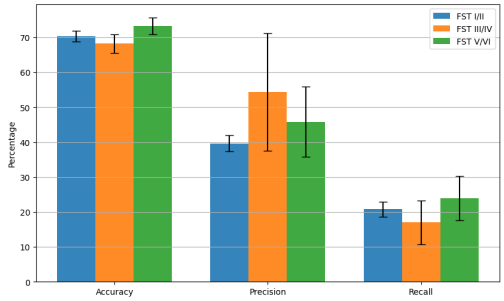

Fig. 8. Claude 3 Opus Malignancy Detection with Zero shot Prompting

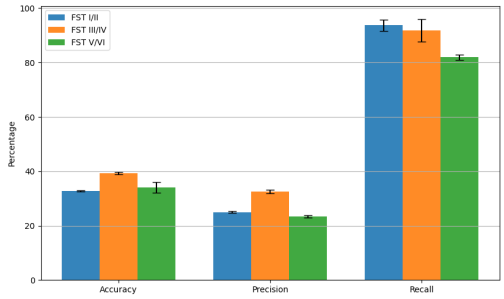

Fig. 9. Claude 3 Opus Malignancy Detection with Few shot Prompting

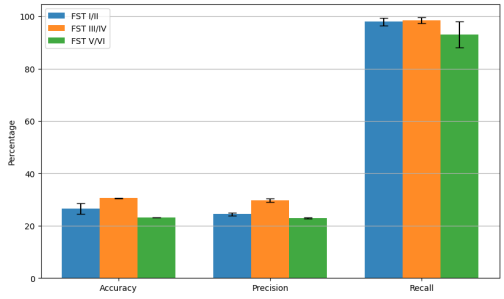

Fig. 10. LLaVA Malignancy Detection with Chain of Thought Prompting