DATA ALIGNMENT FOR ZERO-SHOT CONCEPT GENERATION IN DERMATOLOGY AI

Soham Gadgil *, **Mahtab Bigverdi** *
Paul G. Allen School of Computer Science and Engineering
University of Washington
{sgadgil,mahtab}@cs.washington.edu

ABSTRACT

AI in dermatology is evolving at a rapid pace but the major limitation to training trustworthy classifiers is the scarcity of data with ground-truth concept level labels, which are meta-labels semantically meaningful to humans (Li et al., 2019). Foundation models like CLIP (Radford et al., 2021) providing zero-shot capabilities can help alleviate this challenge by leveraging vast amounts of image-caption pairs available on the internet. CLIP can be fine-tuned using domain specific image-caption pairs to improve classification performance. However, CLIP's pre-training data is not well-aligned with the medical jargon that clinicians use to perform diagnoses. The development of large language models (LLMs) in recent years has led to the possibility of leveraging the expressive nature of these models to generate rich text. Our goal is to use these models to generate caption text that aligns well with both the clinical lexicon and with the natural human language used in CLIP's pre-training data. Starting with captions used for images in PubMed articles (Kim et al., 2023), we extend them by passing the raw captions through an LLM fine-tuned on the field's several textbooks. We find that using captions generated by an expressive fine-tuned LLM like GPT-3.5 improves downstream zero-shot concept classification performance.

1 INTRODUCTION

In dermatology, for performing a diagnosis, dermatologists often use concepts, which refer to a clinical lexicon that is used to describe skin disease findings in the dermoscopic images. For example, Melanoma is often associated with the ABCDE rule including asymmetry, border, color, diameter and evolving (Duarte et al., 2021). Thus, learning these concepts from an image can aid in providing diagnostic explanations and building classifiers which are explainable. However, obtaining these concept labels for dermatology is a difficult and time-consuming task since only well-trained dermatologists can accurately describe skin diseases. There are datasets (Codella et al., 2018; Groh et al., 2021) which have high-quality dermoscopic images, but they are either devoid of manual labels, not inclusive of all concepts, or have very limited samples for some concepts.

There have been many advances in fully-supervised learning for medical image classification spanning multiple domains (Yadav & Jadhav, 2019; Islam et al., 2020; Li et al., 2014). However, the same progress has not been achieved in dermatology image analysis due to limited availability of high-quality images with expert annotations. Recently introduced methods like CLIP provide avenues to perform zero-shot classification without the need of labeled datasets. Prior works like MONET (Kim et al., 2023) leverage image-caption pairs from PubMed articles and medical textbooks to fine-tune CLIP models for dermatology. However, the captions used in these academic sources contain medical terms which are not aligned with the pre-training data of CLIP, which includes image-caption pairs found on the internet. We posit that LLMs like GPT variants can be effectively used to model natural human language. Our contributions include (i) using LLMs for data generation by extending the original captions to align them with CLIP's pre-training data and improve downstream performance on zero-shot concept classification, (ii) demonstrating that these LLMs can be further fine-tuned on the field's textbooks to improve their expressiveness.

*Equal contribuation

2 DATASETS

Textbooks With the advent of LLMs, many open-source and closed-source LLM models pre-trained on vast amounts of open internet text data are available. Although, for a specific task like this work, an improved and more informative text in the dermatology field is required that some of these pre-trained models cannot provide. Therefore, fine-tuning a LLM on the desired text set is a crucial solution to this problem. Dermatology textbooks are a good option for fulfilling this requirement. We chose four books for this purpose: Differential Diagnosis In Dermatology (Ashton & Leppard, 2021), General Dermatology (English, 2007), Top 50 Dermatology Case Studies for Primary Care (Reich et al., 2017), and Handbook of Dermoscopy (Malvehy et al., 2006). We used the text from these textbooks to generate prompt and completion pairs for fine-tuning the LLM models as described in section 3.2.

Evaluation Dataset To evaluate the trained CLIP model for zero-shot concept classification, we used the SKINCON dataset (Daneshjou et al., 2022). SKINCON includes 3230 images from the Fitzpatrick 17k skin disease dataset (Groh et al., 2021), densely annotated with 48 clinical concepts, 22 of which have at least 50 images representing the concept. The concepts used were chosen by two dermatologists considering the clinical descriptor terms used to describe skin lesions, such as "plaque", "scale", and "erosion" to name a few. The list of concepts was based on the clinical lexicon used by dermatologists to describe skin lesions and was developed with consultation of the terms listed in one of the most widely used dermatology textbooks - Dermatology (Bolognia et al., 2012).

3 METHODS

3.1 EXPLORATORY ANALYSIS

For training CLIP, the captions need to be tokenized using the CLIP tokenizer before the contrastive learning procedure. All CLIP models use 77 as the maximum tokenized context length, either padding or truncating the caption if it is below or above that length respectively.

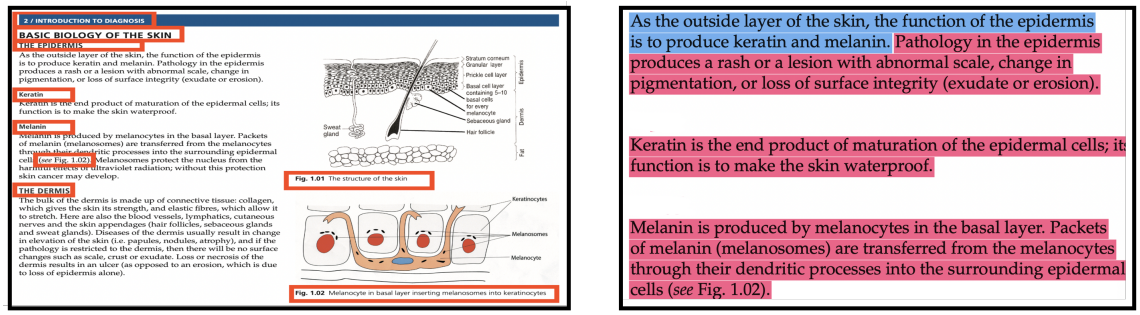
Since we were restricted to 77 as the maximum number of tokens, we first did an exploratory analysis of the tokenized lengths of the original 44314 captions obtained from the PubMed articles utilizing the scripts provided in Kim et al. (2023). This would give us an intuition of how many tokens were available for extending the caption for alignment. Table 2 (Appendix A.2) shows the statistics of the tokenized captions. The mean length of captions is ~ 35 which shows that most of the captions are short and do not exceed the maximum token length of 77. 75% of the captions have a token length of less than 51 which indicates that a majority of captions do have additional tokens available to be extended and improved. There are $\sim 13\%$ captions which have been truncated at the max token length of 77, still leaving around ~ 38000 captions that can be improved using LLMs.

3.2 DATA PREPROCESSING

Fine-tuning data for an LLM needs to be in the form of prompt-completion pairs. It meant for a specific prompt, we needed to define the ideal completion that we expect the model to output. Naively, these would be the sentences that follow a given prompt in the text. However, the whole of the raw text from the books could have misleading phrases and sentences, so applying some preprocessing strategies was essential for fine-tuning data preparation.

We knew that each dermatology book had its own structure, types of references, and formatting method. Preprocessing and extracting a proper text from the books and creating a prompt-completion dataset for further fine-tuning was divided into manual and automatic steps. The manual extraction phase was deleting irrelevant pages like glossaries, acknowledgments, and references. Also, not all text in the preserved pages assisted in creating prompt-completion pairs, such as titles, footnotes, captions, tables' text, and citations. Figure 1 shows some examples. We filtered the main text by picking the lines with the dominant font and size using the PyMuPDF [1] python library. We assumed figures' captions or other non-informative texts like titles are less frequent in the book and have different fonts and sizes. This assumption was valid for all books we used. Table 3 (Appendix A.2) shows the names of the books and the number of prompt-completion pairs obtained for each.

[1] https://github.com/pymupdf/PyMuPDF



a **b**

Figure 1: **a)** Irrelevant and confounding parts of textbooks shown in red boxes are removed from the prompt-completion dataset. **b)** An example of a prompt sentence in blue with the following four sentences in pink as its completion.

3.3 Fine-tuning

We fine-tuned two LLM models, GPT-2 (Radford et al., 2019) and GPT-3.5 (Brown et al., 2020), which have been pre-trained as general purpose learners on a huge amount of text data scraped from the internet. GPT-3.5 is one of the largest autoregressive language models available, trained with 4096-token-long context. However, the model is close-sourced and fine-tuning comes as part of an API endpoint. We first decided to use GPT-2, which is GPT-3.5's predecessor with 1.5 billion parameters. To fine-tune GPT-2, we started with the extracted prompt and completion pairs from the preprocessing step. Then, we created the fine-tuning dataset by combining each prompt and completion into a single sentence separated by the padding token and tokenized the sentence using the GPT-2 tokenizer. Finally, we passed the data to the trainer with the combined prompt and completion as the label. We used the huggingface library (Wolf et al., 2019) to implement the GPT-2 model and fine-tuned it for two epochs. GPT-3.5 was easier to fine-tune and only needed an API key to directly call a fine-tuning endpoint. The `gpt-3.5-turbo` variant of GPT-3.5 was fine-tuned for four epochs using a similar input data from the mentioned books with the format $\{"prompt" : "promptA", "completion" : "completionA"\}$.

For fine-tuning CLIP, we started by extracting the image-caption pairs from PubMed articles using the scripts provided in Kim et al. (2023). We didn't use textbooks here since the repository does not have the list of textbooks used. Then, we passed the captions through the fine-tuned LLM to generated enriched captions with a max length of 512 tokens. Table 4 (Appendix A.2) shows some of the improved captions generated using fine-tuned GPT-2 and GPT-3.5 models. We then fine-tuned the pre-trained CLIP model `openai/clip-vit-base-path32` with a batch size of 64 using the Adam optimizer (Kingma & Ba, 2014) and a learning rate of 1e-5 with a cosine annealing scheduler with warm restarts.

3.4 Zero-shot classification and Evaluation

Once the CLIP model was fine-tuned, we used the 3230 images and corresponding concepts from the SKINCON dataset to perform zero-shot concept classification. For each concept key in the 48 SKINCON concepts, we created embeddings for the text *"This is {concept_key}"* and all of the images in CLIP's joint embedding space. Then, using the cosine similarity scores, we generated a Receiver operating characteristic (ROC) curve independently for each of the 48 concepts. The evaluation metric used was the area under the ROC curve (AUC).

4 Results

We evaluated using five different CLIP models: **1)** The vanilla CLIP model without fine-tuning (Vanilla) and the CLIP model fine tuned using **2)** Original PubMed image-caption pairs (Original).

3) Aligned captions from fine-tuned GPT-2 **4)** Aligned captions from Vanilla GPT-3.5. **5)** Aligned captions from fine-tuned GPT-3.5.

We decided to include vanilla GPT-3.5 in our results since from qualitative analysis it seemed that GPT-3.5 by itself had a high enough expressive power to understand even techincal medical context from the captions and generate customizations. Table 1 shows the mean AUC across all concepts for the different CLIP models as defined above and Table 5 (Appendix A.2) shows the AUC scores for each of the concepts.

Table 1: Mean AUC across all concepts

CLIP Model	Mean AUC
Vanilla	0.572
Original	0.636
Fine-Tuned GPT-2	0.642
Vanilla GPT-3.5	0.639
Fine-Tuned GPT-3.5	**0.648**

From Table 4 (Appendix A.2), it can be seen that the fine-tuned GPT-2 model is able to extend the input caption while keeping the sentence grammatical correct. However, it sometimes strays away from the context of the input caption and can start constructing sentences by stringing together medical jargon. This might be a result of setting a high max token length which causes the model to lose context in longer ranges. GPT-3.5 is able to maintain context for a longer token length and performs better data alignment.

Fine-tuning the CLIP model improves performance for most of the concepts (41 out of 48), see Table 5 (Appendix A.2). The fine-tuned GPT-3.5 model performs the best among all the models tested, with an AUC of 0.648 and it performs better than the original model in a majority of the concepts (26 out of 48). This indicates that fine-tuning the LLM using dermatology text helps in improving the data alignment in the extended captions.

The second best performing model is the GPT-2 fine-tuned model, with an AUC of 0.642 and performing better than the original model in 25 out of the 48 concepts. This result was unexpected since the GPT-3.5 model is much more powerful in terms of the model capacity as compared to GPT-2 and we expected the Vanilla GPT-3.5 model to outperform the fine-tuned GPT-2 model, which was not the case. This indicates that fine-tuning LLM models does actually improve the predictive performance even if the model does not have as many trainable parameters.

The Vanilla GPT-3.5 model is also able to outperform the Original model with an AUC of 0.639. This shows that LLMs can be effectively used to produce customized and well-aligned captions which improve the language supervision provided to the CLIP training procedure resulting in improved performance.

5 Conclusion

Our study reveals that extending captions through the use of a fine-tuned Large Language Model (LLM) on dermatology textbooks effectively connects clinical lexicon with CLIP's pre-training data, resulting in enhanced downstream zero-shot concept classification performance in dermatology images. To summarize, our findings underscore the promise of LLMs in enhancing language supervision for dermatology AI. The improved CLIP model can be further used to annotate images with concepts that can be crucial to developing concept-based disease classifiers like concept bottleneck models (Koh et al., 2020) that are interpretable and transparent. However, further investigation is essential to optimize integration of LLMs with domain-specific models, ensuring more resilient applications in medical image analysis.

6 ACKNOWLEDGMENT

This work was done as part of the final project for the course CSE 527 (Computational Biology) at the University of Washington. We would like to thank the professor Dr. Su-In Lee along with the teaching assistants Wei Qiu and Mingyu Lu for their valuable feedback.

7 LIMITATIONS

Although the early findings are promising, there are many ways to extend this project. We only used 4 dermatology textbooks for extracting the prompt-completion pairs to fine-tune the LLMs, but there are a lot more books available which can also be preprocessed. PubMed articles can be used to generate the prompt-completion pairs as well. Also, in the extraction pipeline, we made pairs by getting the following four sentences of a particular sentence without considering the context and paragraph switch. This could introduce confounders in the fine-tuning process. For instance, the first completion sentence could be related to melanoma; in contrast, the other three could be from the next section and discuss another disease. In addition, python pdf parsers occasionally fail and break some words into meaningless chunks that can doubtlessly mislead the LLM during fine-tuning. A solution for the extraction issues is adding more manual and automatic steps to remove and filter meaningless words and checking the context integration. LLMs have also been known to hallucinate (Lee et al., 2018; Bang et al., 2023) and proper steps need to be taken to ensure non-existent facts are not fabricated which is pertinent in a high-stakes domain like dermatology.

Furthermore, we used the `gpt-3.5-turbo` variant of GPT-3.5, but there are more powerful variants available like GPT-4 which we did not use due to budget constraints. Another approach to enhance the performance of the fine-tuned Large Language Model (LLM) and refine the generated captions is by incorporating Instruction Tuning data (Zhang et al., 2023; Liu et al., 2023; Dai et al., 2023; Ouyang et al., 2022), instruction-output pairs, extracted from dermatology books during the fine-tuning process. This task needs a careful plan to create a dataset that is useful and gives valuable insights.

Another change that could be made is the CLIP model used. We used the `openai/clip-vit-base-path32` model for CLIP training but there is a more powerful baseline CLIP model `openai/clip-vit-large-patch14` available, which we did not use because of memory constraints and longer training times. We can also employ a non-random batch sampling strategy, which includes samples with different concepts in one mini-batch for efficient learning of concepts. Another way to improve language supervision by employing ways to increase the number of tokens from 77, which is CLIP's limitations. We anticipate that all of these changes will improve the zero-shot classification performance of the fine-tuned CLIP model.

REFERENCES

Richard Ashton and Barbara Leppard. *Differential diagnosis in dermatology*. CRC Press, 2021.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

Jean L Bolognia, Joseph L Jorizzo, and Julie V Schaffer. *Dermatology e-book*. Elsevier Health Sciences, 2012.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 168–172. IEEE, 2018.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

Roxana Daneshjou, Mert Yuksekgonul, Zhuo Ran Cai, Roberto A. Novoa, and James Zou. Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=gud0qopqJc4.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ana F Duarte, Bernardo Sousa-Pinto, Luís F Azevedo, Ana M Barros, Susana Puig, Josep Malvehy, Eckart Haneke, and Osvaldo Correia. Clinical abcde rule for early melanoma detection. *European Journal of Dermatology*, 31(6):771–778, 2021.

Mohamed Elhoseiny, Yizhe Zhu, Han Zhang, and Ahmed Elgammal. Link the head to the" beak": Zero shot learning from noisy text description at part precision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5640–5649, 2017.

John SC English. *General Dermatology*. Atlas Medical Publishing Limited, 2007.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1820–1828, 2021.

Md Mohaimenul Islam, Hsuan-Chia Yang, Tahmina Nasrin Poly, Wen-Shan Jian, and Yu-Chuan Jack Li. Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis. *Computer Methods and Programs in Biomedicine*, 191: 105320, 2020.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019.

Chanwoo Kim, Soham U Gadgil, Alex J DeGrave, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee. Fostering transparent medical image ai via an image-text foundation model grounded in medical literature. *medRxiv*, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. Hallucinations in neural machine translation. 2018.

Cheng-Xu Li, Chang-Bing Shen, Ke Xue, Xue Shen, Yan Jing, Zi-Yi Wang, Feng Xu, Ru-Song Meng, Jian-Bin Yu, and Yong Cui. Artificial intelligence in dermatology: past, present, and future, 2019.

Qing Li, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng, and Mei Chen. Medical image classification with convolutional neural network. In *2014 13th international conference on control automation robotics & vision (ICARCV)*, pp. 844–848. IEEE, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

Josep Malvehy, Ralph P Braun, Susana Puig, Ashfaq A Marghoob, and Alfred W Kopf. *Handbook of dermoscopy*. CRC Press, 2006.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

Tzuf Paz-Argaman, Yuval Atzmon, Gal Chechik, and Reut Tsarfaty. Zest: Zero-shot learning from text descriptions using textual similarity and visual summarization. *arXiv preprint arXiv:2010.03276*, 2020.

Sarah Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. *arXiv preprint arXiv:2209.03320*, 2022.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Danya Reich, Corinna Eleni Psomadakis, and Bobby Buka. *Top 50 Dermatology Case Studies for Primary Care*. Springer, 2017.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, pp. 1–8, 2022.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Samir S Yadav and Shivajirao M Jadhav. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, 6(1):1–18, 2019.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction tuning for large language models: A survey, 2023.

A APPENDIX

A.1 RELATED WORK

The CLIP network (Radford et al., 2021) learns visual concepts by being trained with image and text pairs in a self-supervised manner, using text paired with images found across the Internet. CLIP uses a contrastive learning procedure to generate a multi-model embedding space by jointly training an image encoder and a text encoders such that the embeddings of a given image-text pair are close together in the joint representation space. Given a batch of N (image, text) pairs, CLIP is trained to predict which of the $N \times N$ possible (image, text) pairings across a batch actually occurred. This is done by maximizing the cosine similarity of the image and text embeddings of the N real pairs in the batch while minimizing the cosine similarity of the embeddings of the $N^2 - N$ incorrect pairings. This optimization is done using a symmetric cross entropy loss over these similarity scores. CLIP is powerful enough to be used in zero-shot manner on standard images (such as those from ImageNet (Deng et al., 2009) classes). However, dermatology images are sufficiently different from everyday images that it would be useful to fine-tune CLIP with them.

There has been prior work done for performing self-supervised constrastice learning tasks in the medical domain. (Tiu et al., 2022) used contrastive learning for training a self-supervised model on chest x-ray images lacking explicit annotations to perform pathology-classification tasks. However, the MIMIC-CXR dataset (Johnson et al., 2019) which was used to train the model consists of expert radiology reports accompanying each image which has rich textual descriptions about the x-ray and enables the text transformer to better learn visual medical concepts and generalize to different pathologies. In case of dermatology images, no such dataset exists containing images with corresponding expert reports.

In the language supervision domain, several prior works have used text-based knowledge of image categories to improve classification accuracy. Elhoseiny et al. (2017) extracts visual information from unstructured text descriptions collected from the internet to recognize object parts and perform zero-shot classification. Paz-Argaman et al. (2020) extract visual information from Wikipedia descriptions to enable bird classification. These works show that text augmentation is useful for self-supervised models, but they rely on external natural language datasets for generating descriptions. Pratt et al. (2022) uses a large language model (GPT-3.5) to generate customized prompts for a range of zero-shot image classification benchmarks, however it does not perform any fine-tuning and the datasets don't include any images from the medical domain.

Our proposed approach combines the prior works by leveraging the expressive power provided by LLMs with the availability of dermatology images and contrastive learning approaches to improve zero-shot concept generation using customized prompts.

A.2 TABLES

Table 2: Tokenized caption statistics

Statistic	Value
Mean	35.394
Standard Deviation	22.801
Minimum	3
Lower Quartile	17
Median	28
Upper Quartile	51
Maximum	77

Table 3: Dataset statistics

Book Name	Pages	Number of pairs
Differential Diagnosis In Dermatology (Ashton & Leppard, 2021)	466	616
General Dermatology (English, 2007)	147	286
Top 50 Dermatology Case Studies for Primary Care (Reich et al., 2017)	341	851
Handbook of Dermoscopy (Malvehy et al., 2006)	106	58
Total		1811

Table 4: Examples of improved captions using fine-tuned GPT-2 and GPT-3.5

Caption	GPT-2 Fine-Tuned	GPT-3.5 Fine-Tuned
Long-standing disseminated granuloma annulare on the left leg.	Infection with streptococci can cause pustular psoriasis on the lower legs, a condition which is commonly referred to as "sore throat on the legs." Granuloma annulare is a common, self-limiting, skin-tearing condition. It occurs in about 10 % of immunocompromised populations and occurs predominantly at sites where skin is heavily covered, such as the arms and legs.'	This long-standing disseminated granuloma annulare (DGA) on the left leg is likely caused by an unknown infection. The DGA typically presents with a red, inflamed lump that is hard to see. It can cause significant pain and difficulty walking. Treatment typically involves antibiotics and/or surgery.
Anakinra-induced sarcoidosis. Buttock erythematous plaque eruption while on anakinra for hidradenitis suppurativa.	Multiple papillary lesions scattered symmetrically over the lower abdominal area with or without a border, sometimes with or without scaling. Itching on palpation.The clinical appearance and pathogenesis of sarcoid vary greatly depending on the size, location and age of the patient. In the elderly, it is typically chronic, and lasts for many years.	Anakinra is a medication used to treat certain inflammatory conditions, including hidradenitis suppurativa. In rare cases, anakinra has been linked to an increased risk of developing sarcoidosis, a condition that causes abnormal collections of inflammatory cells in the body. Signs and symptoms of sarcoidosis can vary but can include skin rashes and lesions, such as an erythematous plaque eruption on the buttocks.
Interdigital furunculosis with a painful, erythematous and alopecic nodule.	These frequently rupture to drain a haemopurulent discharge. Individual lesions may be often caused by penetrating foreign bodies, but recurrent lesions are more commonly associated with atopic dermatitis and/or conformational problems.The discharge is red, erythematous and itchy. It gets better spontaneously after few days to a week.	Atopic eczema tends to be poorly defined, but with prominent warty plaques. In peridial eczema the surface is dry and erythematous, and is complicated by eosinophilia and by chronic, unresponsive or nodular form.The differential diagnosis of peridial eczema includes solar eczema, rubella and psoriasis may turn from pink to deep purple with the development of scabs.

Table 5: Concept Classification AUC. # True shows the number of samples in the SKINCON dataset labeled with that concept

Concept	# True	Vanilla	Original	Fine-Tuned GPT-2	Vanilla GPT-3.5	Fine-Tuned GPT-3.5
Vesicle	46	0.552	0.667	**0.715**	0.674	0.584
Papule	1170	0.448	0.66	0.646	0.668	**0.676**
Macule	13	0.407	0.395	0.54	0.469	**0.556**
Plaque	1967	0.566	0.646	0.582	**0.656**	0.611
Abscess	5	**0.929**	0.847	0.787	0.875	0.884
Bulla	64	0.508	0.611	0.584	0.647	**0.654**
Patch	149	0.547	0.461	0.546	**0.599**	0.523
Nodule	189	0.719	**0.773**	0.744	0.758	0.758
Ulcer	154	0.82	0.883	**0.886**	0.879	0.883
Crust	497	0.559	0.666	0.635	0.671	**0.727**
Erosion	200	0.538	0.593	**0.626**	0.603	0.602
Excoriation	46	0.536	**0.693**	0.6	0.559	0.578
Atrophy	69	0.482	0.606	0.613	0.563	**0.616**
Exudate	144	**0.677**	0.656	0.617	0.629	0.626
Purpura/Petechiae	10	0.577	0.592	0.662	**0.667**	0.646
Fissure	32	0.708	0.548	0.428	0.506	**0.686**
Induration	33	**0.594**	0.559	0.528	0.573	0.553
Xerosis	35	0.41	0.735	0.737	**0.744**	0.547
Telangiectasia	100	0.366	0.484	**0.574**	0.47	0.564
Scale	686	0.485	0.474	0.434	0.417	**0.521**
Scar	123	0.604	**0.659**	0.592	0.568	0.639
Friable	153	**0.629**	0.576	0.628	0.555	0.377
Sclerosis	27	**0.661**	0.557	0.582	0.595	0.506
Pedunculated	26	0.665	**0.855**	0.755	0.817	0.773
Exophytic/Fungating	42	**0.713**	0.629	0.657	0.607	0.7
Warty/Papillomatous	46	**0.71**	0.591	0.592	0.636	0.691
Dome-shaped	146	0.624	0.604	**0.71**	0.658	0.667
Flat topped	18	0.574	0.595	0.609	0.563	**0.635**
Brown(Hyperpigmentation)	760	0.648	0.768	**0.776**	0.763	0.738
Translucent	16	0.496	0.523	0.69	**0.731**	0.547
White(Hypopigmentation)	257	0.596	0.686	0.718	0.715	**0.737**
Purple	85	0.725	**0.843**	0.813	0.777	0.762
Yellow	245	0.614	**0.744**	0.733	0.706	0.721
Black	90	0.685	0.873	**0.901**	0.882	0.896
Erythema	2139	0.609	**0.719**	0.711	0.666	0.68
Comedo	24	0.469	0.502	0.527	0.561	**0.632**
Lichenification	25	0.505	**0.565**	0.55	0.545	0.502
Blue	5	0.662	0.749	0.767	0.754	**0.784**
Umbilicated	49	0.57	0.683	0.567	0.663	**0.751**
Poikiloderma	5	0.324	**0.621**	0.453	0.4	0.524
Salmon	10	0.463	**0.671**	0.641	0.667	0.588
Wheal	21	0.507	**0.796**	0.775	0.666	0.693
Acuminate	8	0.444	0.279	0.588	**0.654**	0.606
Burrow	5	**0.807**	0.636	0.585	0.68	0.786
Gray	5	0.302	0.45	0.439	0.283	**0.303**
Pigmented	5	0.459	0.483	0.513	0.581	**0.661**
Cyst	6	0.521	0.745	**0.883**	0.79	0.827
Mean AUC		0.572	0.636	0.642	0.639	**0.648**