YuanAI at MEDIQA-MAGIC 2024: Improving Medical VQA Performance through Parameter-Efficient Fine-Tuning

Notebook for the ImageCLEF Lab at CLEF 2024

Hsian-Hong Fu[1,*,†], Hsien-Cheng Huang[1,†]

[1]Yuan Ze University, Taoyuan, Taiwan

Abstract

In our participation in the MEDIQA-MAGIC[1] 2024 workshop at CLEF, we employed the mediqa-m3g-dataset[2] for fine-tuning and one-shot sampling. Our primary models for inference were Llama3 and Gemini, where Gemini served as the Vision-Language Pre-training (VLP) model and Llama3 was utilized for downstream tasks. We focused on parameter-efficient fine-tuning of Llama3 using Low-Rank Adaptation (LoRA). Our approach achieved notable results, including a DELTA-BLEU score of 4.461 and a BERTScore of 0.855, the highest in the task competition. This study underscores the efficacy of parameter-efficient fine-tuning techniques in enhancing medical visual question answering (VQA) performance.

Keywords

Large Language Models(LLM), Large multimodal model(LMM), Llama3, gemini, Parameter-Efficient Fine-Tuning(PEFT),

1. Introduction

In this paper, we focus on describing the techniques used in the MEDIQA-MAGIC[1] task, including pre-processing, fine-tuning, and inferencing. Also, we used the mediqa-m3g-dataset[2] to train, validate, and test our model. The dataset consists of medical questions and images, with the goal of generating medical advice based on the input. To effectively extract information from images without re-training a multimodal model, we opted to use a Vision-Language Pre-Training(VLP) model to obtain image information first. We then combined the medical question with the text we've generated from the images. That way, we only requires to fine-tune the subsequent LLM to meet the needs of our downstream tasks. This approach reduces the GPU memory needed for fine-tuning and increases efficiency.

2. Related Work

2.1. Large Language Models

Since the release of transformers, numerous large language models (LLMs) have been introduced, including classic models like BERT and the GPT[3, 4, 5] series. Starting from GPT-3[5], many models have gained widespread attention for their exceptional performance in few-shot and zero-shot learning scenarios. Among them, the recently released Llama3[6] by Meta has achieved remarkable results across various datasets compared to other open-source LLMs.

2.2. Parameter-Efficient Fine-Tuning(PEFT)

Today's large language models (LLMs) require high-quality and extensive datasets. Therefore, fine-tuning the entire model without additional medical datasets becomes an impractical goal. Partial Enabled

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

†These authors contributed equally.

kevin010411@gmail.com (H. Fu); ryankert01@gmail.com (H. Huang)

0009-0005-6769-2501 (H. Fu); 0009-0005-1117-8845 (H. Huang)

Fine-Tuning (PEFT) can achieve more desirable responses even with a small amount of data. PEFT encompasses various techniques such as "Reparameterized, Additive, Partial and Hybrid Fine-Tuning,"[7] each offering diverse fine-tuning methods.

In Reparameterized Fine-Tuning, notable methods like LoRA[8] introduce additional learnable variables in Linear Layers or Attention layers.

In Additive Fine-Tuning involves fixing LLM parameters and adding learnable parameters in front of the original prompt. Examples include prefix-tuning[9], which trains new prefixes to provide context and guide LLMs to produce appropriate answers, and Parameter-Efficient Prompt Tuning[10], which modifies the input layer context without changing the core model parameters. Adapter-based methods, such as AdapterDrop[11] and Hadamard Adapters[12], introduce adapters to adapt the model to new tasks.

In Hybrid Fine-Tuning methods, such as MAM Adapter[13] and AutoPEFT[14], establish connections between the aforementioned techniques. These methods allow for effective model fine-tuning to adapt to downstream tasks even without access to high-quality and extensive datasets.

2.3. Vision language models

Vision language models encompass a wide range of domains, involving both text and images. These models integrate distinct features from both modalities to perform various downstream tasks, such as automatic subtitle generation for videos and visual question answering (VQA). The vision component of these models can be traced back to the field of image classification, where labels are used to categorize each image, often with multiple labels per image. Backbone networks such as ResNet[15], EfficientNet[16], ViT[17], and Swin[18] are employed to extract features for classification purposes. Subsequently, tasks like image captioning have emerged, where backbone networks extract image features, and use Language model to generate simple text descriptions of the image content. Notable examples of this approach include the CLIP[19] and BLIP series[20, 21]. In the context of our task presented in this paper—MEDIQA-MAGIC[1]—we focus on the VQA domain. Here, we extract features from images or convert them into textual descriptions, which are then processed by large language models (LLMs) to produce the desired descriptions.

Vision Question Answering (VQA) has long been an established field. Historically, models like RCNN or ResNet were used to extract image features, which were then aligned with text using models like BERT or transformers to generate answers, as seen in models like ViLBERT[22], VL-BERT[23], and Pixel-BERT[24].

One mainstream method involves using an image encoder and a text encoder, followed by a transformer to combine the two features. Notable examples of this approach include models like LLaVA[25], BEiTv3[26], and InternVL[27].

3. Methodology

3.1. Overall Approach

In this paper, we describe the techniques we've employed in the MEDIQA-MAGIC[1] task, focusing on data processing, model fine-tuning, and inference. To efficiently extract information from images without re-training a multimodal model, we utilized a Vision-Language Pre-trained (VLP) model to obtain image information. This information was then combined with the text extracted from the images. Subsequently, only the downstream Large Language Model (LLM) required fine-tuning. This approach minimizes GPU memory usage during fine-tuning and enhances overall efficiency. In this task we only ues T4 16GB RAM on Colab to fine-tuning and inference.

In this section, we will provide a detailed explanation of the steps involved, including data preprocessing, model pretraining, model inference and post-processing. As illustrated in figure3.1, we adopted the approach of converting images to text using a VLP model to bridge the gap between images and text. However, VLP models without specific pretraining may not effectively convert medical images.

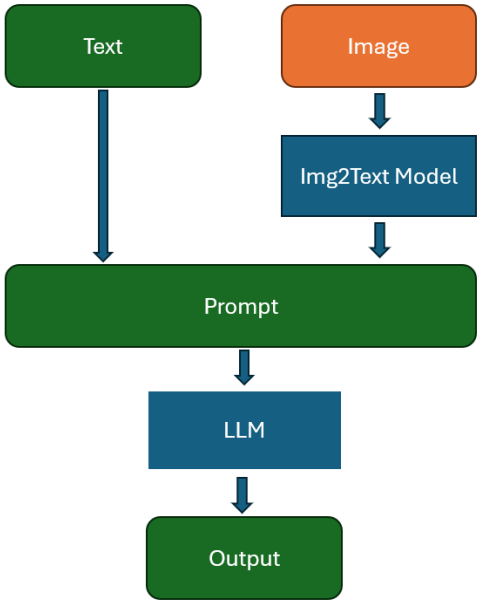

Figure 1: Our workflow visualized

Therefore, without using additional training datasets, we employed the more universally effective Gemini[28] for this role. For the LLM, we used the latest Llama3-8b[6] model.

3.2. Image2Text processing

We used Gemini[28] for data preprocessing to convert images into text. Gemini offers better image descriptions compared to VLP models like CLIP[19] and BLIP[20, 21].

3.3. pretraining on text2text model

Using the LoRA method provided by UnSloth, we incorporated LoRA[8] into Llama3-8b[6] with their default parameters: r=16, dropout=0, alpha=16. Subsequently, we employed SFT (Supervised Fine-Tuning) to optimize LoRA in order to match the style of the correct answers. During fine-tuning, we did not utilize the query_content_en since most of the training data lacks substantial content. Therefore, we only used query_title_en as inputs. Additionally, for all instructions, we uniformly used "Give me a medical advice" as the prompt, then input the patient context inside query_title_en.

3.4. text prompt engineering

Initially, we used a fixed training example for one-shot learning, which meant we selected a single, consistent example from our training data. However, this approach proved ineffective due to the varying suitability of examples for different tasks. To improve this, we switched to randomly selecting an example from our training data for each one-shot attempt. This enhancement allowed us to better match the diversity of tasks and improve the model's adaptability.

3.5. model inference(LLM) & post-processing(Output)

During the inference phase, we not only utilize the query_title_en of the current question but also incorporate zero, one, two, or four shots randomly selected from the training data. We refrain from using more shots due to the limitations imposed by max_length. Initially, during testing, we included the text generated from image conversion in the one-shot input. However, experimental results showed a decrease in performance on validation due to the max_length constraint. Therefore, in the final

Table 1
Data Distribution

Dataset	All Data number	Contain Image
Train	310	271
Validation	50	44
Test	93	93

Table 2
Compare the performance on the test set with and without fine-tuning

Model(deltableu/BERTScore)	Zero Shot	One Shot	Two Shot	Four Shot
No Fine Tune	1.393/0.820	1.462/0.826	0.991/0.820	1.611/0.826
LoRA Fine Tune	2.073/0.829	3.467/0.831	5.522/0.836	6.673/0.835

evaluation, only the text generated from the current question's image conversion is included as part of the input.

After generating the text, leading and trailing whitespace is removed from the output. However, due to the limitation of max_new_tokens, it's not guaranteed that <end_token> will appear in every output. Thus, there's no assurance of complete answers in the actual response every time.

3.6. Evaluation Methods

In Table 1, we observe a total of 310 training samples, out of which 271 include image data. For the validation set, there are 50 samples, with 44 containing images. The test set comprises 93 samples, all of which include images.

Regarding evaluation metrics, we employ two scoring methods: DeltaBLEU[29] and BERTScore[30]. BERTScore[30] leverages contextual embeddings from BERT pre-trained models to compute the F1 score based on the maximum cosine similarity between two sentences, with a maximum score of 100. The actual algorithm for BERTScore is shown in Equation 3. DeltaBLEU[29] Score is derived from BERTScore and incorporates weights derived from human qualitative judgments, as well as the maximum number of N-gram matches across multiple reference answers, yielding a maximum score of 1. The actual algorithm for DeltaBLEU is shown in Equation 4

$$\text{Precision}(\mathbf{X}, \mathbf{Y}) = \frac{1}{m} \sum_{x_i \in \mathbf{X}} \max_{y_j \in \mathbf{Y}} \text{sim}(x_i, y_j) \tag{1}$$

$$\text{Recall}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n} \sum_{y_j \in \mathbf{Y}} \max_{x_i \in \mathbf{X}} \text{sim}(y_j, x_i) \tag{2}$$

$$\text{BertF1Score}(\mathbf{X}, \mathbf{Y}) = 2 \cdot \frac{\text{Precision}(\mathbf{X}, \mathbf{Y}) \cdot \text{Recall}(\mathbf{X}, \mathbf{Y})}{\text{Precision}(\mathbf{X}, \mathbf{Y}) + \text{Recall}(\mathbf{X}, \mathbf{Y})} \tag{3}$$

$$\Delta\text{BLEU} = \frac{\sum_i \sum_{g \in \text{n-grams}(h_i)} \max_{j: g \in r_{i,j}} \{w_{i,j} \cdot \#g(h_i, r_{i,j})\}}{\sum_i \sum_{g \in \text{n-grams}(h_i)} \max_j \{w_{i,j} \cdot \#g(h_i)\}} \tag{4}$$

4. Experiments

4.1. Result

Table 2 shows a significant difference in scores before and after fine-tuning. Without fine-tuning, the performance of one-shot learning heavily depends on the quality of the example used. Since we used random sampling from the training data to assist LLM predictions during one-shot tasks, if the sampled

Table 3
Compare the performance on the validation set with different img2Text model

Image2Text Model(deltableu/BERTScore)	Zero Shot	One Shot	Two Shot	Four Shot
no image Text	5.494/0.848	2.050/0.850	2.998/0.846	2.361/0.844
BLIP	2.202/0.846	6.861/0.847	1.631/0.846	6.376/0.850
Gemini	2.073/0.829	3.467/0.831	5.522/0.836	6.673/0.835

Table 4
BLIP and Gemini Text Generated from the Same Imaged

Image ID	BLIP	Gemini
11mk4th	The image is of a woman's eye	The image shows a close-up of a person's left eye, focusing on the lower eyelid. There appears to be a small, yellow bump or pustule near the inner corner of the eye. The surrounding skin is slightly red and inflamed. The person's eye is open, but their expression and the context of the image (e.g., setting, posture) are unclear. It's difficult to determine the person's age and gender based on the image alone.
11wvv7y	The image is of a small white spot on the tip of a finger	The image shows a close-up of a fingertip with a small, irregular-shaped, white-yellowish lump. The lump appears dry, crumbly, and slightly raised. There are no other visible skin conditions, injuries, or discolorations on the finger or surrounding area. The background suggests a well-lit, clean environment, possibly a medical setting. The focus on the isolated lump suggests concern about its nature and origin. Due to the limited information, it is difficult to determine the patient's age, gender, and overall health condition, as well as the context surrounding the appearance of the lump.
11rk8rd	No Image describe	The image depicts a close-up view of a skin lesion on the back of a person's arm, just below the elbow. The lesion is roughly circular, approximately 1 centimeter in diameter, and raised with a well-defined border. It exhibits a red or pink coloration and has a scaly, whitish surface with some crusting. The surrounding skin appears normal in color and texture. There is no visible hair within the lesion, although hair is present on the surrounding skin. The setting and the individual's age and gender are unclear, as is their expression or posture, making it difficult to assess their overall condition or level of discomfort.

data were of low quality, the scores might not improve even with multiple samples. After applying LoRA fine-tuning, the model showed a trend where more shots generally resulted in higher scores. However, due to the max_token limitation of LLMs, we could not indefinitely add more shots to enhance performance.

Table 4 shows the results of converting randomly selected images into text using both Gemini[28] and BLIP[21]. It is evident that BLIP[21] often fails to successfully generate text descriptions and, when it does, the descriptions tend to be overly brief. In contrast, Gemini[28] consistently uses more detailed text to describe the images. We believe that in the medical domain, providing more detailed descriptions is crucial. Therefore, we conclude that using Gemini is more effective in assisting Llama3[6] in answering questions.

From Table 3 compares the different results obtained from testing img2Text data generated by various img2Text models. we can see that "no image text" performed well in the zero-shot setting. This is because our model was fine-tuned using LoRA without image text, and increasing the number of shots without image text did not yield higher scores. Although BLIP[21] performed well in one-shot and four-shot settings, it did not perform well in the two-shot setting. We believe this is because BLIP often fails to effectively convert images to text, so random sampling does not guarantee better results with more samples. For the results generated by Gemini[28], we observed a clear trend of improved

performance with more shots. We believe this is because Gemini provides more stable and reliable results compared to BLIP[21].

At the end, to give you a better understanding of our method, we provide an example of the input and output in appendix Table 5.

5. Limitations

After analyzing the results, we identified several limitations in our approach. First, when diagnosing, the model often mistake the disease for another similar disease. That way, the model may provide incorrect advice. Although sometimes, the advice is still correct due to the simularity of the diseases. But, more than half of the time, the advice is incorrect and may cause harm to the patient. Therefore, it's irresponsible to use our model in real-world medical applications.

6. Future Work

Our current model has limitations in accurately distinguishing between similar diseases, often leading to incorrect advice despite sometimes being correct due to disease similarities. To address this, we propose a new method for future implementation that combines Chain of Thought (CoT)[31] reasoning with Retrieval-Augmented Generation (RAG)[32].

In this method, we aim to leverage the patient's provided symptoms to narrow down the potential diseases. The Chain of Thought approach will help the model logically deduce and refine the possible diagnoses step by step. By integrating CoT, the model will follow a structured reasoning process, improving its ability to distinguish between similar diseases.

Furthermore, we will combine this with RAG to enhance the model's access to relevant medical information. RAG will allow the model to retrieve and incorporate external medical knowledge dynamically, providing more context and supporting the CoT reasoning process.

By combining these approaches, we expect to improve the accuracy and reliability of our model's medical advice. Future work will involve implementing and testing this combined method, utilizing more advanced models like Llama3 70b[6] and GPT-4-o[33]. With these enhancements, we believe our model will better support healthcare professionals in making informed decisions.

7. Conclusion

This study demonstrated an efficient approach to the MEDIQA-MAGIC task using Llama3 and Gemini models. By fine-tuning Llama3 with LoRA and leveraging Gemini for image-to-text conversion, we significantly improved performance, achieving high DELTA-BLEU metrics. Our results highlight the effectiveness of parameter-efficient fine-tuning methods and the importance of detailed image descriptions in medical AI applications.

However, our observations from the BERTScore and manual review indicate that the model often misidentifies diseases, leading to inaccuracies in the provided answers. This poses a significant limitation in medical question answering or diagnostic assistance, as the accuracy of the information is crucial. Providing inaccurate information could mislead healthcare professionals, potentially causing harm. Therefore, despite the model's fluency in generating sentences, the issue of model hallucination remains unresolved. This necessitates further improvements in the model's ability to accurately identify diseases before it can be considered viable for commercial use in the medical field.

References

[1] W. Yim, A. Ben Abacha, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, Overview of the mediqa-magic task at imageclef 2024: Multimodal and generative telemedicine in dermatology, in: CLEF 2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[2] W. Yim, Y. Fu, Z. Sun, A. Ben Abacha, M. Yetisgen, F. Xia, Dermavqa: A multilingual visual question answering dataset for dermatology, CoRR (2024).

[3] K. N. Alec Radford, Improving language understanding by generative pre-training, 2018. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[4] J. W. Alec Radford, Language models are unsupervised multitask learners, 2019. URL: https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf.

[5] B. M. Tom B. Brown, Language models are few-shot learners, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[6] metaAI Team, Introducing meta llama 3: The most capable openly available llm to date, 2024. URL: https://ai.meta.com/blog/meta-llama-3/.

[7] H. X. Lingling Xu, Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment, 2023. URL: https://arxiv.org/abs/2312.12148.

[8] E. J. Hu, Y. Shen, P. Wallis, Lora: Low-rank adaptation of large language models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[9] P. L. Xiang Lisa Li, Prefix-tuning: Optimizing continuous prompts for generation, 2021. URL: https://arxiv.org/abs/2101.00190.

[10] N. C. Brian Lester, Rami Al-Rfou, The power of scale for parameter-efficient prompt tuning, 2021. URL: https://arxiv.org/abs/2104.08691.

[11] A. Rücklé, G. Geigle, M. Glockner, T. Beck, J. Pfeiffer, N. Reimers, I. Gurevych, AdapterDrop: On the efficiency of adapters in transformers, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7930–7946. URL: https://aclanthology.org/2021.emnlp-main.626. doi:10.18653/v1/2021.emnlp-main.626.

[12] Y. Chen, Q. Fu, G. Fan, L. Du, J.-G. Lou, S. Han, D. Zhang, Z. Li, Y. Xiao, Hadamard adapter: An extreme parameter-efficient adapter tuning method for pre-trained language models, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 276–285. URL: https://doi.org/10.1145/3583780.3614904. doi:10.1145/3583780.3614904.

[13] X. M. Junxian He, Chunting Zhou, Towards a unified view of parameter-efficient transfer learning, 2022. URL: https://arxiv.org/abs/2110.04366.

[14] H. Zhou, X. Wan, I. Vulić, A. Korhonen, Autopeft: Automatic configuration search for parameter-efficient fine-tuning, arXiv preprint arXiv:2301.12132 (2023). URL: https://arxiv.org/abs/2301.12132.

[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. URL: https://arxiv.org/abs/1512.03385. arXiv:1512.03385.

[16] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. URL: https://arxiv.org/abs/1905.11946. arXiv:1905.11946.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL: https://arxiv.org/abs/2010.11929. arXiv:2010.11929.

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, 2021. URL: https://arxiv.org/abs/2103.14030. arXiv:2103.14030.

[19] J. W. K. Alec Radford, Learning transferable visual models from natural language supervision, 2021. URL: https://arxiv.org/abs/2103.00020.

[20] D. L. Junnan Li, Blip: Bootstrapping language-image pre-training for unified vision-language

understanding and generation, 2022. URL: https://arxiv.org/abs/2201.12086.
[21] D. L. Junnan Li, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL: https://arxiv.org/abs/2301.12597.
[22] D. B. Jiasen Lu, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019. URL: https://arxiv.org/abs/1908.02265.
[23] X. Z. Weijie Su, Vl-bert: Pre-training of generic visual-linguistic representations, 2019. URL: https://arxiv.org/abs/1908.08530.
[24] Z. Z. Zhicheng Huang, Pixel-bert: Aligning image pixels with text by deep multi-modal transformers, 2020. URL: https://arxiv.org/abs/2004.00849.
[25] C. L. Haotian Liu, Visual instruction tuning, 2023. URL: https://arxiv.org/abs/2304.08485.
[26] H. B. Wenhui Wang, Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022. URL: https://arxiv.org/abs/2208.10442.
[27] J. W. Zhe Chen, Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2023. URL: https://arxiv.org/abs/2312.14238.
[28] G. Gemini Team, Gemini: A family of highly capable multimodal models, 2023. URL: https://arxiv.org/abs/2312.11805. arXiv:2312.11805.
[29] M. Galley, C. Brockett, A. Sordoni, Y. Ji, M. Auli, C. Quirk, M. Mitchell, J. Gao, B. Dolan, deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets, in: C. Zong, M. Strube (Eds.), Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Beijing, China, 2015, pp. 445–450. URL: https://aclanthology.org/P15-2073. doi:10.3115/v1/P15-2073.
[30] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, 2020. URL: https://arxiv.org/abs/1904.09675. arXiv:1904.09675.
[31] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, Q. Le, D. Zhou, Chain of thought prompting elicits reasoning in large language models, CoRR abs/2201.11903 (2022). URL: https://arxiv.org/abs/2201.11903. arXiv:2201.11903.
[32] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.
[33] OpenAI, J. Achiam, S. Adler, S. Agarwal, Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

A. appendix

A.1. Online Resources

Our sample code for model fine-tuning and inferencing is available online at the following links:

- Training Sample Colab,
- Testing Sample Colab.

Table 5
An Example of Input and Output of our method

title	text
User text input	Some group of people has a genetic predisposition to these lines. It is more common in the dominant arm. Research studies are limited on this condition but the more fat that you carry on your arms, the more likely creases are to form on your skin. If you don't want this crease, then first lose some body fat to rule out that cause. If the crease still persists, then you'll know that they're a permanent fixture on your arms due to your genetics. If they're not causing you any pain or if they don't look too abnormal for your liking , then you don't need to worry about them.
Gemini img2text result	The image shows a close-up of the back of a person's hairy leg. The individual's age and gender are indeterminate from the image. The skin appears to have a slightly reddish hue and several small, raised bumps. These bumps could be insect bites, folliculitis, or another type of skin irritation. There is no visible evidence of injury, swelling, or medical equipment in the image. The setting and context of the image are unclear, as it only shows a close-up of the person's leg. More information about the individual's symptoms and medical history would be needed to make a definitive diagnosis.
Output	It is a case of eczema called dyshidrotic eczema. It is aggravated by atopic dermatitis, excessive sweating, sun exposure, and smoking. Avoid aggravating factors where possible. Potassium permanganate soaks may be useful in the acute phase. Apply topical treatments like topical steroids (usually potent or ultrapotent), pimecrolimus and tacrolimus, and regular use of emollients and moisturizers. I have also added a cream for faster healing to be applied twice daily for 2 weeks. Use white toothpaste. Avoid direct contact with food items. Take multivitamins once daily. Rest and I/V fluids are must. Antibiotics are given in case of systemic infection. The lesion can be surgically removed. Referral to a dermatologist is recommended for dermoscopic examination and skin biopsy. The 5-year survival rate is >95%.