Multimodal MoE for InfographicsVQA

Stanford CS224N Custom Project

Manolo Alvarez
Department of Electrical Engineering
Stanford University
manoloac@stanford.edu

Abstract

Most engineers waste hours a day digging for multi-modal information. Aimed at saving precious time, this project is designed to leverage and enhance question-answering capabilities of Large Vision Language Models (LVLMs) in complex technical contexts. Through a Mixture of Experts (MoE) framework, this project attempts to address common computational inefficiencies in multi-modal learning without compromising on results. The agent, an MoE LLaVA model, is tuned via Quantized Low-Rank Adapters (QLoRA), maximum-likelihood training on the InfographicsVQA datasets (SFT), and implicit reward models through Direct Preference Optimization (DPO). In evaluation, the tuned model significantly improves the performance of the baseline in MM-Bench & MM-Vet benchmarks, and in qualitative studies.

1 Key Information to include

- Mentor: Soumya Chatterjee
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

Engineers spend significant time searching for multi-modal information across dense technical documents, graphical presentations, and numerical data — a task both time-consuming and prone to errors that has the potential to evolve significantly with the advent of Large Vision Language Models (LVLMs). Current solutions, while general in application, fall short in addressing the unique challenges posed by the engineering field, which demands a blend of precision, speed, and the ability to reason over complex, multi-modal datasets. This project, therefore, seeks to fill this gap by introducing the first LVLM specifically fine-tuned for engineering reasoning and tasks that leverages a Mixture of Experts (MoE) framework for computational efficiency.

The approach builds on the foundational MoE-LLaVA model (Lin et al., 2024), a promising model that, despite its impressive capabilities, often produces results that are either slightly inaccurate or computationally demanding for engineering applications. By adapting this model through Supervised Fine-tuning (SFT) and Direct Preference Optimization (DPO) on the InfographicsVQA dataset, I aim to create a more efficient, precise, and contextually aware AI agent. Results suggest that this specialized LVLM can indeed far surpass baseline performances in technical contexts, highlighting its potential to revolutionize how engineers interact with information, reduce search times, and make informed decisions more rapidly.

3 Related Work

The landscape of Large Vision Language Models (LVLMs) has been rapidly evolving, with researchers striving to balance the computational cost against the performance benefits these models offer. A notable stride in this direction is the MoE-LLaVA model, which introduces a Mixture of Experts (MoE) framework to LVLMs. A MoE approach selectively activates the top-k experts - feed-forward neural networks - through routers during deployment, leaving other experts inactive. This framework is designed to tackle the inefficiencies inherent in multi-modal learning by selectively activating subsets of the model, thus enabling a sparse, yet powerful, architecture that maintains computational efficiency without compromising on model performance. The MoE-LLaVA model represents a significant leap towards addressing the prohibitive training and inference costs associated with scaling LVLMs, providing a promising foundation for future explorations into efficient multi-modal learning.

The MoE-LLaVA model's approach to model sparsity and efficiency is not without precedents; however, it refines and builds upon the concept of dynamic routing within models, which has been explored in various contexts. Previous efforts in this area have largely focused on the general application of MoE frameworks to large language models, aiming to improve computational efficiency or model performance across a range of tasks. The work of Noam Shazeer (2017) on introducing MoE into deep neural networks for language tasks offers foundational insights into this approach, demonstrating the potential for significant performance gains. Yet, upon personal analysis, these earlier implementations have not fully addressed the unique challenges of multi-modal data processing in technical and engineering contexts, where the integration of textual and visual information demands both precision and efficiency.

This project differentiates itself by focusing on the application of the MoE framework within the realm of engineering, aiming to harness the strengths of LVLMs specifically for engineering reasoning and tasks. By fine-tuning the model on the InfographicsVQA dataset presented in Minesh Mathew (2021) through maximum-likelihood, and aligning it to human preferences via Direct Preference Optimization (DPO) (Rafael Rafailov, 2023), the project not only seeks to enhance its question-answering capabilities but also to ensure that the answers provided are more intuitively aligned with the expectations and preferences of engineering professionals. This deliberate integration of DPO into the model's fine-tuning process represents an innovative step towards creating AI systems that are not just technically proficient but also contextually and practically relevant, offering insights into the future of personalized, efficient, and effective multi-modal learning systems tailored to specific domain needs.

4 Approach

The project employs a combination of quantized low-rank adapters (QLoRA) (Tim Dettmers, 2023), Supervised Fine-tuning (SFT), DPO, and MoE-Tuning to align the model to the task. The **baseline** MoE-Llava model uses the Phi-2 2.7 billion parameter transformer by Microsoft and the CLIP model - a ViT-L/14 Transformer image encoder with a masked self-attention Transformer as a text encoder - by OpenAI.

4.1 LoRA

In LoRA 1 (Edward J. Hu, 2021), the weights of the neural network are modified by adding a low-rank matrix. If we consider a weight matrix W in the attention layers of transformer, the LoRA modification is applied as $W' = W + BA$, where W' is the modified weight matrix. B and A are the low-rank matrices with a rank factor of r. This means if W is of dimension $d \times k$, then B is of dimension $d \times r$ and A is of dimension $r \times k$.

During training, only the parameters of the low-rank matrices B and A are updated. The original parameters of the base model remain frozen. To apply the LoRA modification, we add the BA product to the original weight matrices of the self-attention layers Q, K, and V.

QLoRA Tim Dettmers (2023), 4-bit quantizes the model's weights on top of LoRA 1, shrinking the model size by a factor of 4 at the expense of parameter precision. In practice, quantization can slightly hinder model performance, but it was a conscious trade-off that was made in order to fit the models within personal hardware.

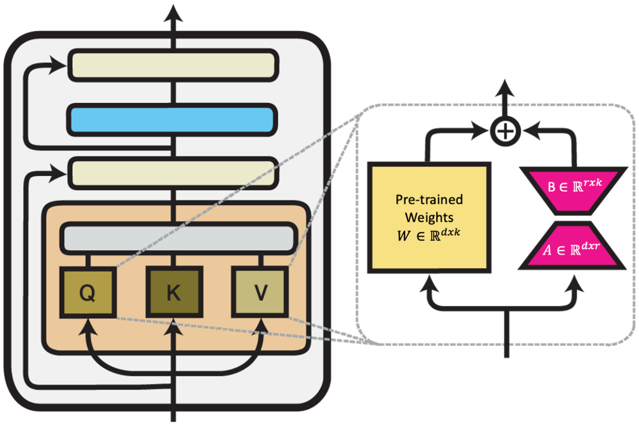


Figure 1: Low-Rank Adapters

4.2 DPO

The traditional approach in optimizing human preferences through reinforcement learning (RL) involves using an auxiliary reward model to fine-tune the main model, encouraging it to produce more high-reward outputs and fewer low-reward ones. The Direct Policy Optimization (DPO) method streamlines this process by directly optimizing the language model using preference data. It achieves this through a novel analytical approach that converts the RL loss involving both reward and reference models into a loss solely based on the reference model, eliminating the need for a complex RL-based optimization process.

To understand DPO, it is helpful to analyze the gradient of the loss function with respect to the parameters θ:

$$\nabla_\theta L_{DPO}(\pi_\theta;\pi_{ref}) =$$

$$-\beta\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\underbrace{\sigma(\hat{r}_\theta(x,y_l)-\hat{r}_\theta(x,y_w))}_{\textit{higher when reward estimate is wrong}}\left[\underbrace{\nabla_\theta\log\pi(y_w|x)}_{\textit{increase likelihood of } y_w} - \underbrace{\nabla_\theta\log\pi(y_l|x)}_{\textit{decrease likelihood of } y_l}\right]\right] \tag{1}$$

where $\hat{r}_\theta(x,y) = \beta\log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}$ is the reward implicitly defined by the language model π_θ and reference model $\pi_{ref}(y|x)$. As the labels in equation 1, note, the gradient of the loss function will increase the likelihood of the preferred completions, decrease the likelihood of the dis-preferred completions, and scale the update by how much higher the implicit reward model $\hat{r}_\theta$ rates the dis-preferred completions Rafael Rafailov (2023).

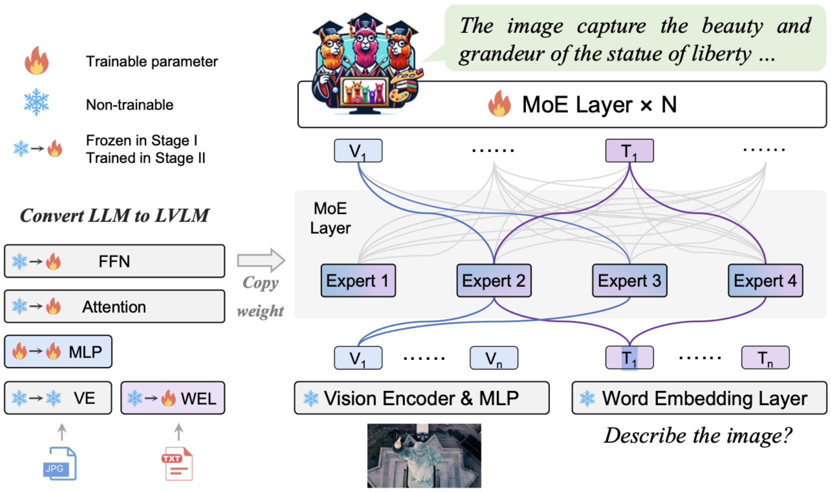
4.3 MoE-Tuning

The architecture of MoE-LLaVA, depicted in 2 , comprises several key components: a vision encoder, a visual projection layer (MLP), word embedding layer, multiple stacked LLM blocks, and MoE blocks. Upon **SFT**, the model undergoes a three-stage MoE-Tuning/training strategy, presented in Lin et al. (2024):

1. Train the MLP to adapt the LLM to visual inputs.
2. Freeze the MLP weights and train the LLM without involving MoE layers.
3. Freeze the MLP and non-FFN weights, replicate the weights of the FFN to instantiate the different experts, and train.

4.4 Multimodal MoE for InfographicsVQA

Now that we've covered each technique separately, let's walk through my method for the task, as depicted in figure 3, step-by-step:



(a) Stage I and Stage II

(b) Stage III

Figure 2: **Illustration of MoE-Tuning.** The MoE-Tuning consists of three stages. In stage I, only the MLP is trained. In stage II, all parameters are trained except for the Vision Encoder (VE). In stage III, FFNs are used to initialize the experts in MoE, and only the MoE layers are trained. For each MoE layer, only two experts are activated for each token, while the other experts remain silent.

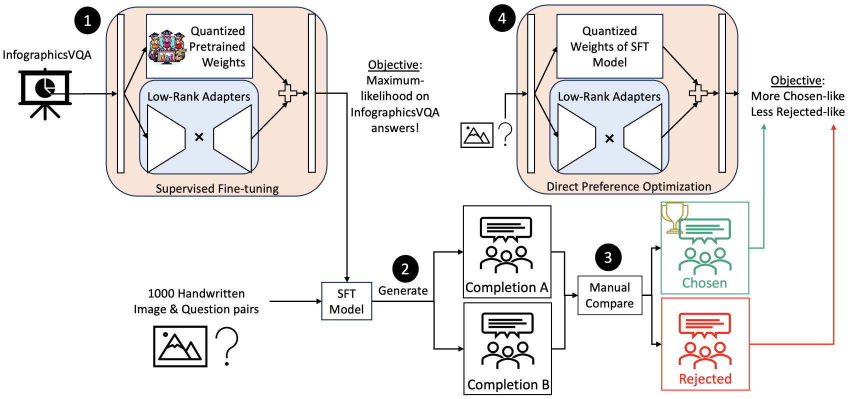

1
InfographicsVQA
Quantized Pretrained Weights
Low-Rank Adapters
×
Supervised Fine-tuning
Objective: Maximum-likelihood on InfographicsVQA answers!
4
Quantized Weights of SFT Model
Low-Rank Adapters
×
Direct Preference Optimization
Objective: More Chosen-like Less Rejected-like
1000 Handwritten Image & Question pairs
SFT Model
2
Generate
Completion A
Completion B
3
Manual Compare
Chosen
Rejected

Figure 3: **Method**

1. **Train** the baseline model with QLoRA to maximize the likelihood of generating the responses in the InfographicsVQA dataset given the corresponding image and prompt pairs. This is the Supervised Fine-tuning (SFT) step.

2. **Generate** completion pairs on 1000 handwritten visual questions using the SFT model.

3. **Annotate** the "chosen" and "rejected" response for each completion pair.

4. **Train** the base model with QLoRA, starting with the trained adapters from the SFT step, via DPO. The objective is to maximize the likelihood of the model generating the "chosen" responses and decrease the likelihood of it generating the "rejected" responses, given its corresponding prompt.

For this project, the MoE-Llava codebase referenced in Lin et al. (2024) was leveraged for MoE training but it was significantly adapted for training in 4-bit and with the InfographicsVQA dataset. No prior code existed for DPO with MoE LVLMs so I had to create a unique DPOTrainer and miscellaneous code that could train the MoE LVLM for the objective.

5 Experiments

5.1 Data

The dataset used for SFT is called InfographicsVQA and it is a collection of 3288 Q&A pairs on 579 infographics compiled by Minesh Mathew (2021). The questions in it require joint reasoning over the document layout, textual content, graphical elements, and data visualizations. They are curated to necessitate elementary reasoning and basic arithmetic skills. Perfect for the aforementioned downstream task.

The dataset used for DPO, as mentioned in the approach, is my collection of hand-annotated preference pairs on 1000 handwritten visual questions answered by the SFT model.

5.2 Evaluation method

Quantitative evaluation of the model was performed on the following benchmarks:

- MM-Bench - approximately 3000 questions on 4378 images spanning 20 ability dimensions. Each question is in a multiple-choice format with a single correct answer. (Yuan Liu, 2023)
- MM-Vet - a collection of 187 images from various online sources with 205 q&a pairs, each of which requires one or more capabilities (out of 6) to answer. Questions are varied in type and entail open-ended responses of differing lengths. (Weihao Yu, 2023)
- Llava-Bench - a diverse set of 24 images with 60 questions in total, including indoor and outdoor scenes, memes, paintings, sketches, etc, and associated with each image, a highly-detailed and manually-curated description. `https://github.com/haotian-liu/LLaVA/blob/main/docs/LLaVA_Bench.md`

MM-Bench and MM-Vet are more arithmetically demanding benchmarks; well suited for this application. In contrast, Llava-Bench is a collection of not so relevant VQA's that I believed important to include to understand general performance of the model on other tasks.

Qualitative evaluation was performed across various image & question pairs in Electrical, Software, and Thermal Engineering domains. Visualizations of these can be observed in section 6.

5.3 Experimental Details

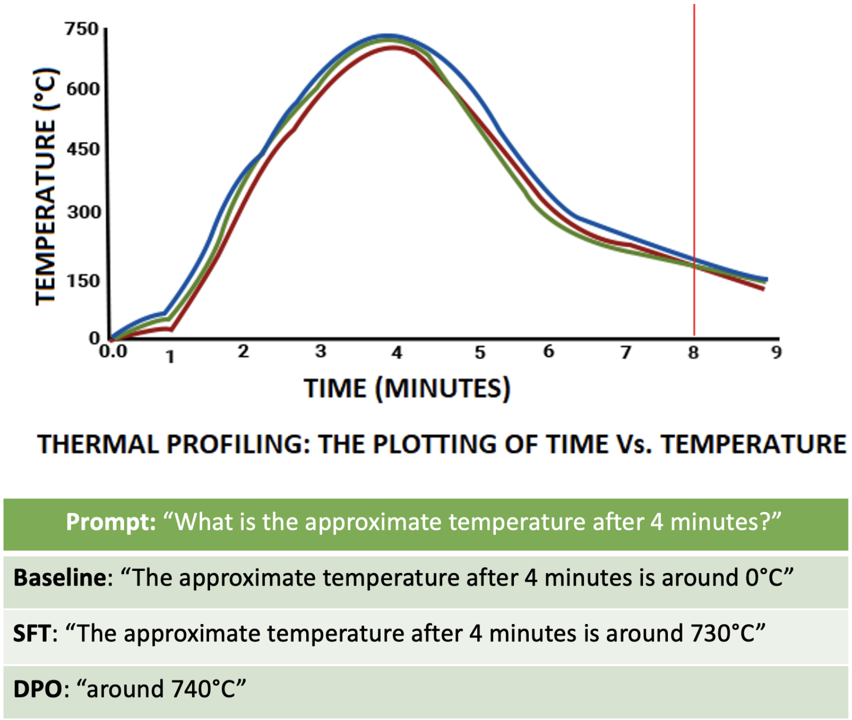
- Number of Experts = 4
- Top-K Experts = 2
- Train Epochs = 3
- Train Batch Size = 4
- Gradient Accumulation Steps = 2
- Lora Ranks = 8
- Lora Alpha = 16
- Lora Dropout = 0.05
- Max Sequence Length = 512
- Optimizer = Adamw Torch
- Quant Type = nf4
- Learning Rate = 2e-5
- LR Scheduler Type = cosine
- Training Time = about 5 hours
- Hardware = x1 RTX 4090 16GB GPU

5.4 Results

As can be seen in table 1, the model fine-tuned (SFT) on the InfographicsVQA dataset improved the performance of the baseline on the more arithmetically demanding benchmarks, MM-Bench and MM-Vet, and worsened on generic image q&a benchmarks like Llava-bench, as hoped. It is important to highlight that the MM-Vet score nearly doubled that of the baseline. The DPO model, as expected, closely matched the SFT scores. For reference, I also included the scores of other popular LVLM's in the table. In section 6, I analyze these results quantitatively and qualitatively.

Table 1: Baseline vs Fine-tuned Model Benchmark Results

Model	MM-Bench	MM-Vet	Llava-Bench
Baseline	65.2	34.3	94.1
SFT	68.7	66.2	73.8
DPO	67.6	64.9	75.5
GPT-4V	N/A	67.7	N/A
Gemini Pro Vision	73.6	64.3	N/A



Prompt: “What is the approximate temperature after 4 minutes?”

Baseline: “The approximate temperature after 4 minutes is around 0°C”

SFT: “The approximate temperature after 4 minutes is around 730°C”

DPO: “around 740°C”

Figure 4: **Completion** for question on a thermal graph.

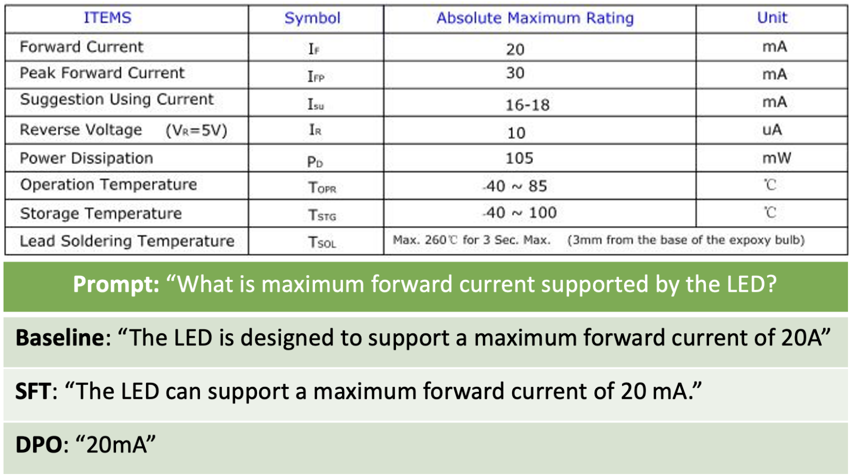
6 Analysis

Improvement over the baseline was expected for the dataset-like benchmarks but the leap on MM-Vet was beyond expectations. I did a qualitative analysis on the MM-Vet completions and while I did prefer the completions of the tuned model, I am unsure they deserve double the score. A more detailed analysis across several hundred samples may be merited for future work.

The DPO model matching the SFT scores was expected and desired. My intent was to align the model with human preferences without degrading accuracy. As can be seen in figures 4 and 5, the DPO step aligned the SFT model towards more concise answers. Matching my preferences in the DPO dataset. This is important for me because most LVLM’s I’ve interactive with are wordy (repeat the question) unless you instruct it not to be. Forcing the user to read through way more text than necessary.

In figures 6 and 7, we can see the expert loads through the SFT model for LLaVA Bench and MM-Bench, respectively. The expectation was that the model, through training, would’ve gradually become sparser such that an expert loading pattern emerged. Where some experts are predominant in the first few layers and gradually withdraw from the workload. Unfortunately, per the visuals, the experts don’t seem to have learned any pattern that allows them to divide their tasks in a specific manner. Making the usefulness of the MoE debatable.

Furthermore, I show the distribution of modalities across different experts in figures 8 and 9, respectively. Similarly, the experts don’t seem to develop their own preferences. One can see that the routing distributions for text and image are highly dissimilar. As a result, it isn’t clear that each expert



ITEMS	Symbol	Absolute Maximum Rating	Unit
Forward Current	I_F	20	mA
Peak Forward Current	I_{FP}	30	mA
Suggestion Using Current	I_{su}	16-18	mA
Reverse Voltage (VR=5V)	I_R	10	uA
Power Dissipation	P_D	105	mW
Operation Temperature	T_{OPR}	-40 ~ 85	℃
Storage Temperature	T_{STG}	-40 ~ 100	℃
Lead Soldering Temperature	T_{SOL}	Max. 260℃ for 3 Sec. Max. (3mm from the base of the expoxy bulb)	

Prompt: “What is maximum forward current supported by the LED?

Baseline: “The LED is designed to support a maximum forward current of 20A”

SFT: “The LED can support a maximum forward current of 20 mA.”

DPO: “20mA”

Figure 5: **Completion** for question on an electrical datasheet.

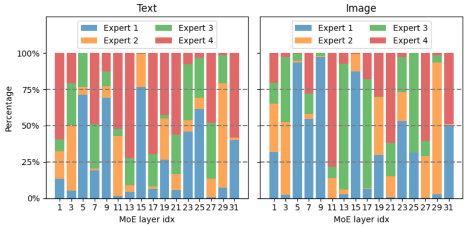


Figure 6: Distribution of modalities across different experts for **LLaVA Bench.**

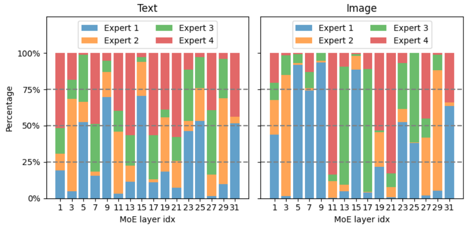


Figure 7: Distribution of modalities across different experts for **MM Bench.**

is capable of handling both text tokens and image tokens simultaneously - something we’d like in an MoE model where only the top-k experts are activated.

Now, clearly the fine-tuned model was well adapted for our task, so my theory is that step 3 in the MoE Tuning process - where only the MoE layers are trained - was not carried out long enough for the experts to learn anything useful. Unfortunately, the original MoE-LLava paper Lin et al. (2024) does not document the MoE layer tuning parameters that made their MoE layers work well, so this is an area that could be further explored.

7 Conclusion

This project demonstrated the feasibility and potential of leveraging and enhancing question-answering capabilities of Large Vision Language Models (LVLMs) in complex technical contexts with consumer hardware. Employing Quantized Low-Rank Adapters (QLoRA), training for maxi-

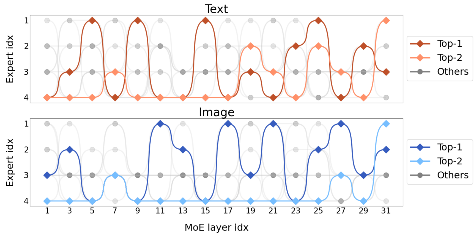


Figure 8: Visualization of activated pathways for **LLaVA Bench.**

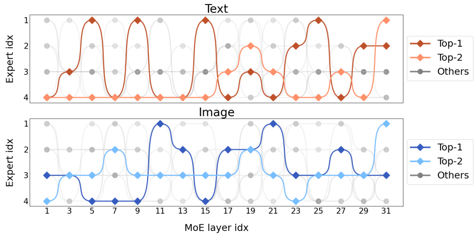


Figure 9: Visualization of activated pathways for **MM Bench.**

mum likelihood (SFT) on the InfographicsVQA dataset, and training with implicit reward models through Direct Preference Optimization (DPO) resulted in models that aligned more closely with human preferences in technical tasks. The key findings include:

1. **Effectiveness of Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO)**: A short bout of SFT can significantly improve a foundational model's ability to generate more precise technical answers. This improvement was evident both qualitatively and quantitatively, indicating the model's enhanced capacity to understand and execute the specific task of visual Q&A in technical contexts. Similarly, a short bout of DPO aligned the model to more concise answers, as preferred.
2. **Challenges with MoE** Contrary to expectations, the MoE layers did not appear to learn anything useful. This was likely attributed to short adaptation bouts of the MoE layers. Future work should explore the training parameters of the MoE layers.
3. **Potential for Further Improvement** Despite the limitations in time and resources, the project's outcomes were promising, showcasing the potential of AI in automating complex tasks like teechnical visual Q&A. Future work with more refined datasets and training time could yield even better results.

In conclusion, this project underscores the significant potential of AI for technical VQA on the edge. While there are challenges to overcome, particularly in dataset generation and training efficiency, the progress made in this project serves as a foundation for future advancements in this domain.

References

Phillip Wallis Zeyuan Allen-Zhu Yuanzhi Li Shean Wang Lu Wang Weizhu Chen Edward J. Hu, Yelong Shen. 2021. Lora: Low-rank adaptation of large language models.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. Online. arXiv.

Rubèn Pérez Tito Dimosthenis Karatzas-Ernest Valveny C.V Jawahar Minesh Mathew, Viraj Bagal. 2021. Infographicvqa.

Krzysztof Maziarz Andy Davis Quoc Le Geoffrey Hinton Jeff Dean Noam Shazeer, Azalia Mirhoseini. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer.

Eric Mitchell Stefano Ermon Christopher D. Manning Chelsea Finn Rafael Rafailov, Archit Sharma. 2023. Direct preference optimization: Your language model is secretly a reward model.

Ari Holtzman Luke Zettlemoyer Tim Dettmers, Artidoro Pagnoni. 2023. Qlora: Efficient finetuning of quantized llms.

Linjie Li Jianfeng Wang Kevin Lin Zicheng Liu Xinchao Wang Lijuan Wang Weihao Yu, Zhengyuan Yang. 2023. Mm-vet.

Yuanhan Zhang Bo Li Songyang Zhang Wangbo Zhao Yike Yuan Jiaqi Wang Conghui He Ziwei Liu Kai Chen Dahua Lin Yuan Liu, Haodong Duan. 2023. Mm-bench.