Efficient Alignment of Medical Language Models using Direct Preference Optimization

Brendan P. Murphy
bigsur@stanford.edu

Introduction

Recent advances in large language models (LLMs) have shown impressive performance on various natural language tasks. However, adapting these models to specialized domains, such as medicine, requires alignment with domain-specific preferences. This research explores the application of Direct Preference Optimization (DPO) in combination with parameter-efficient fine-tuning methods to align a medical LLM, BioMistral-7B, with the nuanced preferences and analytical style required for medical reasoning tasks.

Datasets

PubMed QA
The data for this experiment comes from a subset of the PubMedQA dataset, consisting of 5000 examples of unique medical questions derived from research article abstracts. The dataset is preprocessed by extracting the 'context', 'question', and 'answer' fields and prefixing the 'context' to the 'question'.

Question	Are routine preoperative restaging CTs after neoadjuvant chemoradiation for locally advanced rectal cancer low yield?
Context	Pre-operative restaging CT scans are often performed routinely following neoadjuvant chemoradiotherapy for locally advanced rectal cancer...
Answer	Because of the financial costs and established risks, it may be advisable to take a more selective approach to preoperative imaging.

Evaluation Metrics

- **Human Evaluation: "Win Rate"** Blind study involving a physician as a human evaluator who compares outputs of the Supervised Fine-Tuned (SFT) model and the DPO-optimized model. Evaluator selects preferred output for each question based on understanding of PubMed QA dataset and professional judgment. Win rate calculated as the number of instances where DPO model's output is selected as preferred, divided by total judgments made.
- **DPO Reward Margin** Tracked during DPO training phase to assess model's learning progress. Represents the difference between rewards assigned to preferred and rejected answers. Expected to increase as DPO progresses, indicating model's learning to assign higher rewards to preferred answers. Serves as a proxy for model's accuracy in aligning with preferences defined by DPO dataset.
- **Helpfulness Quotient** Assesses effectiveness of DPO approach in reducing model's tendency to generate evasive or safe responses. Quantifies the extent to which model provides direct and relevant responses to given questions. Calculated using a sample of 500 questions from PubMedQA holdout set and responses generated by unmodified BioMistral-7B, SFT, and DPO-optimized models. Analyzes generated responses to identify instances of evasive or non-specific answers. HQ calculated as:

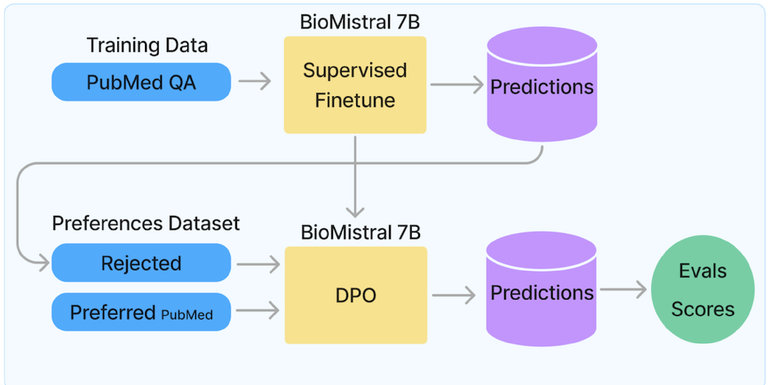
$$HQ = \frac{N_{total} - N_{evasive}}{N_{total}} \quad (1)$$

where N_{total} is the total number of generated responses and $N_{evasive}$ is the number of identified evasive responses.

Methods

The approach involves three main steps:

1. **Supervised Fine-Tuning:** Fine-tune the BioMistral model on the PubMedQA dataset.
2. **DPO Dataset Creation:** Create a DPO dataset by treating the actual answers as preferred choices and the SFT model's predictions as rejected choices.
3. **DPO Optimization:** Apply DPO to the SFT model using the DPO dataset to align the model's outputs with the preferences and style of the dataset.



Supervised Fine-tuning

To efficiently fine-tune BioMistral 7B, I employed these strategies:

- Used the Hugging Face Transformers library and TRL's SFTTrainer
- Employed LoRA for parameter-efficient fine-tuning with a rank of 64, alpha of 16, and dropout of 0.1
- Fine-tuned for 1 epoch with a batch size of 4 per device and a learning rate of 2e-5

Direct Preference Optimization

To optimize my SFT BioMistral 7B model with DPO:

- Loaded the 4-bit quantized version of BioMistral-7B with SFT weights
- Applied LoRA for parameter-efficient DPO tuning with rank 128, alpha 128, and dropout 0.05
- Used TRL's DPOTrainer for optimization with sigmoid loss and beta set to 0.01
- Trained for 1 epoch with a batch size of 8 per device and a learning rate of 5e-6

Results

- DPO-optimized model achieved a 63% win rate over the SFT model in human evaluations.
- Helpfulness Quotient (HQ) improved from 0.81 (unmodified) to 1.00 (DPO-optimized).
- DPO Reward Margin climbed significantly and plateaued around 0.80 epochs, indicating successful alignment with desired preferences.

Model	Win Rate	HQ	Reward Margin
Unmodified BioMistral-7B	-	0.81	-
SFT Model	0.37	0.78	-
DPO Model	**0.63**	**1.00**	**11**

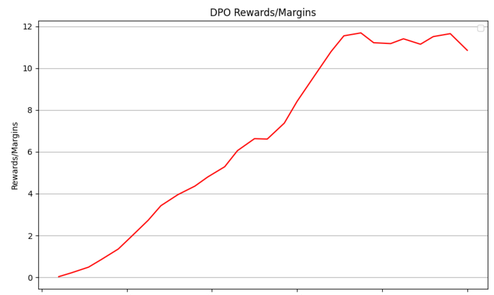


Figure 1. DPO Reward Margin shows the difference between the preferred and rejected answers increasing during training as the model learns the preferred answers have a higher reward than the rejected.

Ancillary Experiment - Summarization

An additional experiment was conducted on the effectiveness of the DPO approach in a medical question summarization task setting using the MeQSum dataset. This showed DPO could effectively align the model to new tasks, even with limited training data, further supporting the findings of the main study.

Model	ROUGE-1			ROUGE-2			ROUGE-L		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
SFT Model	0.12	**0.46**	0.08	0.03	0.12	0.02	0.10	**0.36**	0.06
DPO Model	**0.36**	0.37	**0.38**	**0.23**	**0.23**	**0.24**	**0.34**	0.35	**0.36**

Conclusion

- The proposed approach of combining parameter-efficient fine-tuning with DPO effectively aligns medical LLMs with domain-specific preferences.
- DPO-optimized model outperforms the SFT model in human evaluations and automatic evaluation metrics.
- The Helpfulness Quotient metric quantifies the improvement in generating direct and relevant responses.
- DPO offers a promising direction for developing accurate, relevant, and informative language models in the biomedical field.