Efficient Alignment of Medical Language Models using Direct Preference Optimization

Stanford CS224N Custom Project

Brendan P. Murphy
Department of Computer Science
Stanford University
bigsur@stanford.edu

Abstract

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 efficient fine-tuning techniques and 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 advanced medical reasoning tasks. By leveraging Low-Rank Adaptation (LoRA) and 4-bit quantization, the model is efficiently fine-tuned on the PubMedQA dataset for medical question answering. The supervised fine-tuned model (SFT) is further aligned using DPO, where the ground truth answers from the dataset are treated as preferred outputs, and the SFT model's predictions are considered rejected outputs. Qualitative and quantitative evaluation metrics, including Win Rate, Helpfulness Quotient and DPO Reward Margin, demonstrate the effectiveness of this approach in improving the model's alignment with medical preferences. The DPO-aligned model achieves a 63% win rate over the SFT model in human evaluations and notably eliminates evasive or safe responses that directly avoid answering the question. This research highlights the potential of DPO as a valuable tool for aligning LLMs in specialized domains, offering a more direct and efficient approach to preference learning. The proposed framework of combining parameter-efficient fine-tuning with DPO opens up new possibilities for developing domain-specific LLMs that generate outputs aligned with expert preferences.

1 Introduction

Large language models (LLMs) have demonstrated remarkable performance across a wide range of natural language tasks, driven by the transformer architecture and the increasing scale of data and compute(Vaswani et al., 2023). Models like GPT-3 have shown impressive capabilities in language understanding and generation(Brown et al., 2020).

Aligning language model behavior is important for realizing usefulness and impact of these technologies. However, typical alignment methods pose challenges to adoption due to their complexity and high computational costs. Direct Preference Optimization (DPO) offers a promising alternative, providing a more direct approach to preference learning that avoids the challenges associated with reinforcement learning techniques(Rafailov et al., 2023).

This research explores the application of DPO for aligning a medical language model, BioMistral, for complex clinical reasoning tasks and preferences(Labrak et al., 2024). Unlike common applications of DPO, such as aligning models to be harmless, this project focuses on adapting the model to capture the specific preference and style of the medical domain experts who created the dataset.

To investigate the effectiveness of DPO within a complex clinical reasoning question answering domain. I first performed supervised fine-tuning (SFT) on the BioMistral model using a task-specific dataset. The SFT model served as a benchmark for evaluating the performance of the DPO-optimized model. In the DPO step, I used the same dataset as in the SFT phase, but I treat the actual answers from the dataset as the preferred choices and the SFT model's predicted answers as the rejected choices. By doing so, I aim to align the BioMistral model's outputs with the preferences and analytical style inherent in the dataset. This approach aims to capture nuances, terminology, and reasoning patterns specific to question answering in the medical domain.

To efficiently adapt the BioMistral model while reducing computational resources, the parameter-efficient fine-tuning technique Low-Rank Adaptation (LoRA) was used (Hu et al., 2021). This method freezes most of the model's parameters and fine-tunes only a small subset, significantly reducing training time and memory requirements while maintaining performance and preventing catastrophic forgetting.

This experiment demonstrates promising results, with the DPO-optimized models achieving a win rate of approximately 63% over the SFT model in human evaluations. Additionally, the DPO model no longer evades questions by providing safe responses, indicating improved alignment with the preferred answers. The key contributions of this research are twofold. First, I showcase the effectiveness of DPO in aligning a medical language model to specific tasks and preferences, enabling the generation of outputs tailored to the needs of medical experts. Second, I demonstrate the feasibility of efficiently adapting large language models using parameter-efficient fine-tuning techniques, making the alignment process more accessible and cost-effective.

These findings highlight the potential of DPO as a valuable tool for aligning language models in specialized domains, such as healthcare, where capturing domain-specific preferences and styles is crucial. By providing a more direct and efficient approach to preference learning, DPO opens up new possibilities for developing language models that generate outputs aligned with the needs and expectations of domain experts.

2 Related Work

In the medical domain, domain-specific language models have been developed by pretraining on large amounts of medical text. MEDITRON-70B is a notable example, which scales medical pretraining to a 70 billion parameter model, demonstrating strong performance on various medical language tasks(Chen et al., 2023). These medical LLMs provide a foundation for further adaptation and alignment to specific medical tasks and preferences. Adapting large language models to specific domains and tasks can be achieved through various techniques. Supervised fine-tuning is a common approach, where the model is trained on a labeled dataset specific to the target task. However, fine-tuning large models can be computationally expensive and may require significant resources.

To address the challenges of efficient adaptation, techniques like parameter-efficient fine-tuning have been proposed. Low-Rank Adaptation (LoRA) is a popular method that freezes most of the model parameters and only fine-tunes a small subset, significantly reducing the computational requirements while maintaining performance. LoRA has been successfully applied to medical language models for tasks like question answering and summarization(Wang et al., 2023) and (Van Veen et al., 2024).

Another approach to aligning language models with specific preferences is through Direct Preference Optimization (DPO). DPO involves training the model on a dataset of human-annotated comparisons between model outputs, allowing the model to learn and align with the preferred responses. This technique has been applied to medical text summarization, demonstrating its ability to enhance downstream tasks (Ahn et al., 2024).

Recent work has also explored the use of contrasting responses across identical and diverse prompts to enhance LLM alignment (Yin et al., 2024). They propose a method to optimize the model's preferences by comparing its outputs across different prompts, allowing for a more nuanced alignment with desired behaviors.

Despite the advancements in LLM adaptation and alignment, the evaluation of clinical LLMs for real-world applications remains a critical challenge. Other work introduces the Conversational Reasoning Assessment Framework for Testing in Medicine (CRAFT-MD), a novel approach for evaluating the diagnostic capabilities of clinical LLMs in the context of natural doctor-patient dialogues(Johri et al.,

2024). The authors applied CRAFT-MD to assess GPT-4 and GPT-3.5 in the domain of skin diseases, revealing limitations in their conversational reasoning, history taking, and diagnostic accuracy. Based on these findings, the paper proposes a set of guidelines for future evaluations of clinical LLMs, emphasizing realistic conversations, comprehensive history taking, open-ended questioning, and a combination of automated and expert evaluations.

3 Approach

Direct Preference Optimization (DPO) is an algorithm for training language models from human preferences, without needing explicit reward modeling or reinforcement learning(Rafailov et al., 2023). Prior methods for preference learning fit a reward model on human judgments about model outputs, and then use reinforcement learning to optimize a policy to maximize the predicted reward. However, reinforcement learning is complex and unstable. DPO sidesteps these challenges by parametrizing the reward in terms of the policy directly. Specifically, DPO sets the reward equal to:

$$r(x, y) = \beta \log \left(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} \right)$$

Where π is the policy, π_{ref} is a reference policy, and β controls the strength of the KL constraint that prevents uncontrolled distribution shift. The authors prove this allows transforming the probability of human preferences under a Bradley-Terry model into a loss over policies:

$$L(\pi) = -\mathbb{E}(x, y^+) \log \sigma \left(r(x, y^+) - r(x, y^-) \right)$$

This loss maximizes the relative probability of preferred outputs y^+ over dispreferred outputs y^-. Weighting by the logistic helps prevent instability, so DPO can directly optimize a policy to satisfy human judgments, without needing separate reward modeling or reinforcement learning.

To efficiently adapt the BioMistral language model to specific medical tasks and preferences, this experiment combines DPO with parameter efficient fine tuning techniques. I utilize the Transformer Reinforcement Learning (TRL) library from HuggingFace for supervised fine-tuning and DPO (Wolf et al., 2020) and employ LoRA parameter efficient fine tuning. LoRA is a technique that adds adapters to the existing model, enabling efficient learning and adaptation to downstream tasks while significantly reducing computational resources. By freezing most of the pre-trained model's parameters and introducing a small set of trainable low-rank update matrices, LoRA allows the model to adapt to new tasks while preserving the knowledge acquired during pre-training, thus preventing catastrophic forgetting.

This approach involves three main steps for each medical task:

- Supervised Fine-Tuning (SFT): Fine-Tune the BioMistral model on task-specific datasets and use SFT outputs to establish benchmark performance.

- DPO Dataset Creation: Creating a DPO dataset by treating the actual answers from the task-specific datasets as preferred choices and the SFT model's predictions as rejected choices.

- DPO Optimization: Applying DPO to the SFT model using the DPO dataset, this aligns the models' outputs with the preferences and style of the datasets.

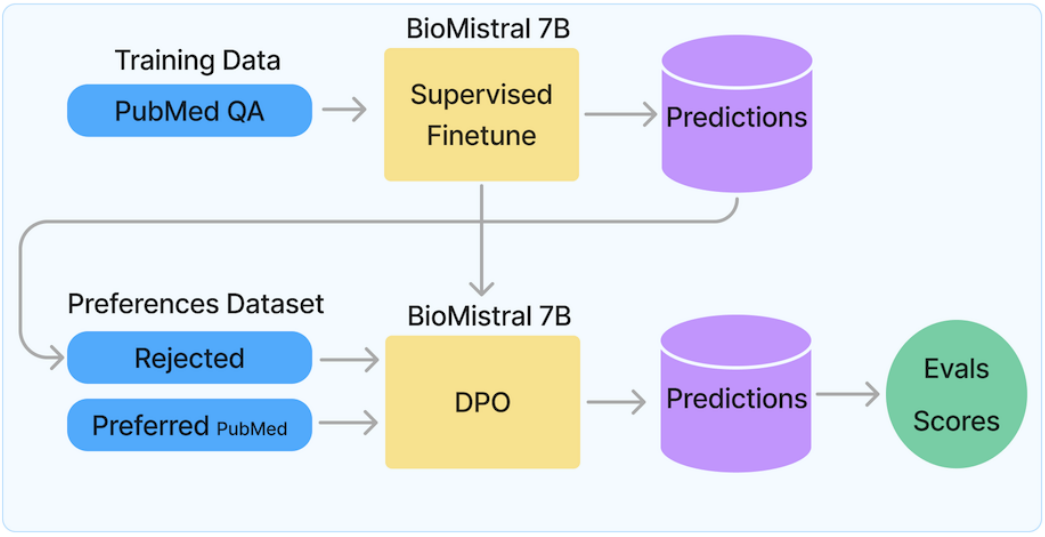

Figure 1: Supervised fine-tuning and DPO of Mistral7B model on PubMed QA Task

4 Experiments

4.1 Data

The data for this experiment comes from a subset of the PubMedQA dataset, a publicly available resource designed for the development and evaluation of medical question answering systems (Jin et al., 2019). This subset consists of 5000 examples selected from PubMedQA, with each example corresponding to a unique medical question derived from the abstracts of research articles. The preprocessing step involves condensing the nested JSON structure of the original PubMedQA dataset and mapping the data into a format suitable for the model. This entails extracting the 'context', 'question' and 'answer' fields, and prefixing the 'context' to the 'question'.

4.2 Experimental Details

Experiments were conducted using the open-source 7 billion parameter BioMistral-7B model on a single NVIDIA A100 GPU with 80GB memory,leveraging NVIDIA CUDA libraries for accelerated computing(NVIDIA et al., 2024). I use the Hugging Face Transformers library's model quantization functionality to reduce the precision of BioMistral-7B's parameters to 4 bits using bitsandbytes which significantly reduces the memory footprint of the model. I employ LoRA to minimize computational overhead and configured training with a rank of 64, alpha of 16, and dropout of 0.1. The target modules for adaptation are the query, key, value and output projection matrices in the attention blocks. I use the TRL library's SFTTrainer for supervised fine-tuning. I also use gradient accumulation with a batch size of 4 per device to simulate larger batch training, and a learning rate of $2e{-}5$ The learning rate schedule follows a cosine decay. I fine-tune for 1 epoch.

5,000 examples from the PubMedQA dataset for question answering in the biomedical domain are used for the SFT step. I preprocess the dataset to map it to a format suitable for prompted fine-tuning, with each example containing a sequence of interleaved user and assistant messages. The dataset is split into 90% train, 5% validation, and 5% test. I limit the total sequence length to the model's maximum of 2048 tokens.

Following the supervised fine-tuning of BioMistral-7B on the PubMedQA dataset, I created the DPO dataset by first generating 5000 predictions from the SFT model on a held out set of PubMedQA questions. For each question, the model's predicted answer is paired with the actual ground truth answer. This forms a dataset where each example consists of a question prompt, the model's prediction (considered the "rejected" answer), and the human-written ground truth (considered the "chosen" answer). I split this dataset into 90% train and 10% test. The dataset is also transformed into the

format expected by the DPO trainer, with each example containing the question prompt, chosen answer, and rejected answer.

For the DPO training, I continue using the 4-bit quantized version of BioMistral-7B, loading the weights fine-tuned during the SFT phase. I apply LoRA on top of this model for parameter-efficient DPO tuning, using a configuration with rank 128, alpha 128, and dropout 0.05. The target modules for LoRA include the query, key, value, output, gate, up, and down projection matrices. I use the TRL library's DPOTrainer for optimization. The training is run for 1 epoch with a batch size of 8 per device, gradient accumulation steps of 4, and a learning rate of 5e-6 with cosine decay. The loss function used is the sigmoid loss, and the DPO hyperparameter beta is set to 0.01. Evaluation is performed every 100 steps on the test split of the DPO dataset. The hypothesis is that the DPO-tuned model will show improved alignment with human preferences compared to the SFT model, generating answers more similar to the human-written ground truth when prompted with questions.

Finally I sample 500 questions from a hold out set from PubMedQA, and use the DPO model to generate long text answers. Automated metrics like training and evaluation loss, Rewards Margin, and Helpfulness Quotient are calculated, as well as human evaluation via inspection by a volunteer rater are used to quantify model performance.

4.3 Evaluation method

To comprehensively assess the performance of the BioMistral-7B model and the effectiveness of the DPO approach, I employ a combination of human evaluation, DPO-specific metrics, and automatic evaluation metrics.

- **Human Evaluation - Win Rate** I conduct a blind study involving a physician as a human evaluator to compare the outputs of the Supervised Fine-Tuned (SFT) model and the DPO-optimized model. The evaluator is presented with a set of questions and the corresponding outputs from both models, without knowing which model generated each output. The evaluator then selects the preferred output for each question based on their understanding of the PubMed QA dataset and professional judgment. To calculate the "win rate," I sample 100 comparisons between the SFT and DPO model outputs. The human evaluator provides judgement by selecting their preferred answer. The win rate is calculated as the number of instances where the DPO model's output is selected as the preferred choice, divided by the total number of judgments made. This metric quantifies the relative performance of the DPO model compared to the SFT model based on human preference.
- **DPO Reward Margin** During the DPO training phase, I track the Reward Margin metric to assess the model's learning progress. The Reward Margin represents the difference between the rewards assigned to the preferred answers and the rejected answers. As the DPO progresses, the Reward Margin is expected to increase, indicating that the model is learning to assign higher rewards to the preferred answers compared to the rejected answers. The Reward Margin serves as a proxy for the model's accuracy in aligning with the preferences defined by the DPO dataset. A higher Reward Margin suggests that the model is effectively distinguishing between preferred and rejected answers, and is adapting its outputs accordingly.
- **Helpfulness Quotient** To assess the effectiveness of the DPO approach in reducing the model's tendency to generate evasive or safe responses, I introduce the Helpfulness Quotient (HQ). This metric quantifies the extent to which the model provides direct and relevant responses to the given questions. I sample a set of 500 questions from the PubMedQA holdout set and generate responses using the unmodified BioMistral-7B model, the Supervised Fine-Tuned (SFT) model, and the DPO-optimized model. I then analyze the generated responses to identify instances of evasive or non-specific answers that do not directly address the given question. The HQ is calculated for each model using the following formula:

$$HQ = \frac{N_{total} - N_{evasive}}{N_{total}} \tag{1}$$

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

By comparing the HQ scores of the unmodified BioMistral-7B model, the SFT model, and the DPO-optimized model, I aim to quantify the extent to which the DPO approach has successfully re-aligned the model to provide more direct and specific answers, reducing its reliance on safe or evasive responses. A higher HQ score in the DPO-optimized model compared to the unmodified and SFT models would indicate the effectiveness of the DPO approach in mitigating this limitation and improving the model's ability to provide relevant and informative answers according to the task-specific training data. The Helpfulness Quotient serves as a targeted metric to evaluate the model's progress in reducing evasive responses and enhancing its overall helpfulness in providing answers that directly address the given questions. By incorporating this metric alongside other evaluation methods, I aim to comprehensively assess the impact of the DPO approach on the model's performance and its alignment with the desired behavior defined by the task-specific training data.

By combining Win Rate, Reward Margin, and Helpfulness Quotient, I provide a comprehensive evaluation of the BioMistral-7B model's performance and the effectiveness of the DPO approach in aligning the model's outputs with the preferred answers. The human evaluation captures the subjective preferences of a domain expert, while the Reward Margin and Helpfulness Quotient offer quantitative measures of the model's learning progress and usefulness, respectively.

5 Results

The performance of the BioMistral-7B model was evaluated using three key metrics: the human evaluator win rate, the Helpfulness Quotient (HQ), and the DPO Reward Margin. Table 1 presents a comparison of the results obtained for the unmodified BioMistral-7B model, the Supervised Fine-Tuned (SFT) model, and the DPO-optimized model.

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

Table 1: Comparison of evaluation metrics for different models

The Win Rate was found to be 63% in favor of the DPO-optimized model, indicating a clear preference for the outputs generated by this model over those of the SFT model.

The Helpfulness Quotient (HQ) metric, which quantifies the proportion of direct and relevant responses, showed a notable improvement from the unmodified BioMistral-7B model (0.81) to the DPO-optimized model (100%). This suggests that the DPO approach has successfully reduced the model's tendency to generate evasive or non-specific answers, resulting in more helpful and informative responses.

The DPO Reward Margin, a metric specific to the DPO training process, measures the difference between the rewards assigned to preferred and rejected answers. During training, the Reward Margin climbed significantly and plateaued around 0.80 epochs. This indicates that the model has effectively learned to distinguish between desirable and undesirable outputs, assigning higher rewards to the preferred answers compared to the rejected ones. The Reward Margin can be interpreted as a measure of the model's accuracy in aligning with the preferences defined by the DPO dataset.

These results demonstrate the effectiveness of the DPO approach in aligning the model's outputs with the preferences of the domain expert and improving the overall helpfulness of the generated responses. The improvement in the HQ metric, in particular, highlights the model's ability to provide more direct and relevant answers after undergoing the DPO optimization process.

The obtained results are highly encouraging, the significant improvement in the human evaluation (win rate), the HQ metric, and the high plateau of the DPO Reward Margin suggests that the DPO approach is a promising method for fine-tuning language models in the biomedical domain. The success of this approach can be attributed to its ability to incorporate expert preferences directly into the model's training process, allowing it to learn and adapt to the desired behavior more effectively than traditional fine-tuning methods. These findings have important implications for the development

of language models in specialized domains, such as biomedicine, where the ability to provide accurate, relevant, and informative responses is crucial. The DPO approach offers a powerful tool for aligning language models with expert preferences and enhancing their performance in real-world applications.

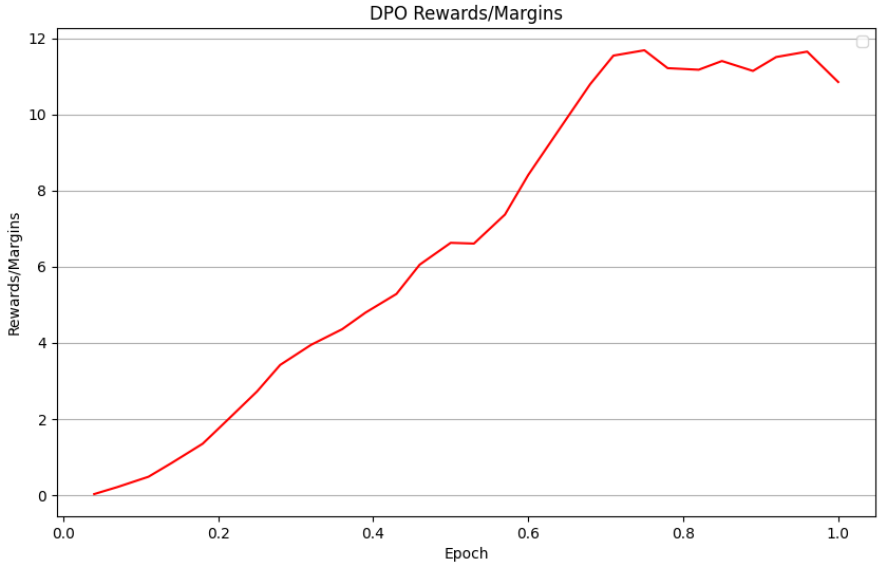

Figure 2: DPO Reward Margin on BioMistral7B and PubMed QA Task. The reward margin shows the difference between the preferred and rejected answers and should increase during training as the model learns the chosen or preferred answers have a higher reward than the rejected answers. This metric can be interpreted as the accuracy.

6 Analysis

The qualitative analysis of the BioMistral-7B model's outputs reveals interesting insights into its behavior and the impact of the DPO approach. The unmodified BioMistral-7B model exhibits a tendency to generate evasive or non-specific answers, particularly when dealing with questions related to subjectivity, race, or ethnicity. This behavior can be attributed to the model's inherent preference for censorship, which was likely acquired during a fine-tuning step after its initial pre-training.

In contrast, the PubMedQA dataset, which serves as the basis for the DPO optimization process, inherently prefers providing the best possible answer to medical questions. By aligning the model's outputs with the preferences of this dataset, the DPO approach effectively reduces BioMistral-7B's censorship preference and encourages the generation of more direct and relevant responses.

The Helpfulness Quotient (HQ) metric, introduced to quantify the proportion of direct and relevant responses, provides a clear indication of the model's qualitative improvement. The notable increase in the HQ score from the unmodified BioMistral-7B model (0.81) to the DPO-optimized model (100), highlights the effectiveness of the DPO approach in reducing evasive or non-specific answers. This improvement in the model's ability to provide helpful and informative responses is a direct result of the alignment with the PubMedQA dataset's preferences.

The human evaluator's preference for the outputs generated by the DPO-optimized model over those of the SFT model further supports the qualitative improvement achieved through the DPO approach. The domain expert's assessment of the model's outputs provides valuable insights into the relevance and informativeness of the generated responses, confirming the effectiveness of the DPO optimization process.

The analysis of the DPO Reward Margin's behavior during training offers additional evidence of the model's successful alignment with the desired preferences. The significant climb and plateau of the Reward Margin at 0.80 epochs indicate that the model has learned to distinguish between

preferred and rejected answers, assigning higher rewards to the desired outputs. This learning process demonstrates the model's ability to capture and internalize the preferences defined by the DPO dataset.

Overall, the qualitative analysis of the BioMistral-7B model's outputs and behavior highlights the power of the DPO approach in capturing and aligning the model with domain-specific preferences. By reducing censorship tendencies and promoting the generation of direct and relevant responses, the DPO-optimized model achieves a significant improvement in its ability to provide helpful and informative answers to medical questions. This analysis underscores the potential of the DPO approach as a valuable tool for adapting language models to specialized domains, such as biomedicine, where accurate and relevant information is of utmost importance.

7 Conclusion

This research demonstrates the effectiveness of combining parameter-efficient fine-tuning techniques, such as LoRA and 4-bit quantization, with DPO to align a medical language model with domain-specific preferences. The proposed approach achieves promising results, with the DPO-optimized model outperforming the SFT model in both human evaluations and automatic evaluation metrics.

The introduction of the Helpfulness Quotient metric provides a valuable tool for quantifying the proportion of direct and relevant responses generated by the model. The significant improvement in the HQ score from the unmodified BioMistral-7B model to the DPO-optimized model highlights the effectiveness of the DPO approach in reducing evasive or non-specific answers and promoting the generation of helpful and informative responses.

The human evaluator's preference for the outputs generated by the DPO-optimized model over those of the SFT model further validates the qualitative improvement achieved through the DPO approach. The domain expert's assessment confirms the relevance and informativeness of the generated responses, underscoring the successful alignment of the model with the PubMedQA dataset's preferences.

The analysis of the DPO Reward Margin's behavior during training provides additional evidence of the model's successful internalization of the desired preferences. The significant climb and plateau of the Reward Margin demonstrate the model's ability to distinguish between preferred and rejected answers, assigning higher rewards to the desired outputs.

While the results of this research are highly encouraging, there are some limitations to consider. Further evaluation of the model's generalization ability on diverse medical datasets and real-world scenarios is necessary to fully assess its performance and robustness. Having multiple evaluators would also provide a more comprehensive and diverse assessment of the model's performance and mitigate subjectivity. Future work should involve a larger pool of evaluators. Additionally, exploring newer modified DPO implementations, such as iterative or online DPO, and alternative preference learning methods that eliminate the need for binary preferences could potentially enhance the model's alignment and adaptability in the medical domain (Ethayarajh et al., 2024).

Overall, this research contributes to the growing body of work on efficient adaptation and alignment of LLMs in specialized domains. The successful application of parameter-efficient fine-tuning techniques and DPO to align a medical LLM with domain-specific preferences opens up new possibilities for the development of accurate, relevant, and informative language models in the biomedical field. The proposed approach offers a promising direction for future research and practical applications, potentially leading to more effective and trustworthy AI-assisted tools in healthcare and medical decision-making.

8 Code

Part of the code for supervised-fine-tuning and dpo was adapted and inspired from a deeplearning.ai notebook, "Supervised fine-tuning (SFT) of an LLM" and "Human preference fine-tuning using direct preference optimization (DPO) of an LLM" by Lewis Tunstall and Edward Beeching of Hugging Face (Tunstall and Beeching, 2024b) and (Tunstall and Beeching, 2024a).

The code is available on my GitHub
https://github.com/csbrendan/cs224N

References

Imjin Ahn, Hansle Gwon, Young-Hak Kim, Tae Joon Jun, and Sanghyun Park. 2024. Note: Notable generation of patient text summaries through efficient approach based on direct preference optimization.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering.

Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I. Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2024. Guidelines for rigorous evaluation of clinical llms for conversational reasoning.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains.

NVIDIA, Péter Vingelmann, and Frank H.P. Fitzek. 2024. Cuda, release: 12.

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

Lewis Tunstall and Edward Beeching. 2024a. Direct preference optimization method notebook. `https://colab.research.google.com/drive/1mWiOFBy3zY6OdINEvHN9EPoQ_VIvfFKw`. Accessed: [Insert Access Date Here].

Lewis Tunstall and Edward Beeching. 2024b. Hugging face notebooks. `https://colab.research.google.com/drive/1WNSVtM82oknmzL1QrJlNu--yNaWbp6o9`.

D. Van Veen, C. Van Uden, L. Blankemeier, J. B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerová, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C. P. Langlotz, J. Hom, S. Gatidis, J. Pauly, and A. S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*. Epub ahead of print.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023. Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Yueqin Yin, Zhendong Wang, Yi Gu, Hai Huang, Weizhu Chen, and Mingyuan Zhou. 2024. Relative Preference Optimization: Enhancing LLM Alignment through Contrasting Responses across Identical and Diverse Prompts.

A Appendix

A.1 Human Study

To calculate the win rate, I sampled 100 SFT to DPO comparisons for which the rater produced 100 judgements. The win rate was calculated as the number of DPO generations selected divided by the total judgements made. 63 / 100 or 63%.

A.2 Participant

I had one volunteer rater, Agustina Saenz (MD, MPH) who blindly compared the benchmark SFT answer to the DPO answer for 100 questions from the PubMed QA dataset. She is a practicing MD and post graduate research at the Rajpurkar Lab at Harvard Medical School as well as a mentor in the Stanford & Harvard Medical AI bootcamp.

A.3 Summarization Experiment

In addition to the main experiment, I conducted a similar experiment for a summarization task utilizing the MeQSum dataset (Ben Abacha and Demner-Fushman, 2019). Although the dataset was likely too small at 1000 examples for the SFT step, the results demonstrate the effectiveness of DPO in redirecting the model to a new task, even with limited data. The SFT model struggled to adapt to the summarization task, often attempting to answer the question instead of providing a summary. This behavior suggests that the SFT step did not have sufficient data to train the model to change its goal from answering questions to summarizing them.

However, when performing DPO with a small hold out set from MeQSum, the model was instructed to summarize the question instead of trying to answer it in most cases. The ROUGE scores presented in Table 2 indicate the significant impact of DPO in redirecting the model to an entirely new type of task.

These results highlight the potential of DPO to effectively align language models to new tasks, even with limited training data. The ability of DPO to redirect the model's focus from question answering to summarization, despite the small dataset size from MeQSum, underscores its versatility and efficiency in adapting to different tasks and objectives.

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**

Table 2: Comparison of ROUGE scores for SFT and DPO models