Robust DPO with Convex-NN on Single GPU

Stanford CS224N Custom Project

Miria Feng
Department of Electrical Engineering
Stanford University
miria0@stanford.edu

Abstract

Fine-tuning language models (LMs) to better align to human preferences is a fast-moving and still emerging area of research. Recently, models such as ChatGPT(OpenAI) and Claude(Anthropic) are able to better assist our daily lives in numerous ways. These models are typically trained via Reinforcement Learning from Human Feedback (RLHF), which is non-trivial to implement. Therefore, Direct Preference Optimization (DPO) emerged as a performant alternative via reparametrization of the objective. This made aligning LMs to human preference significantly more accessible on modest GPU clusters. In this project, we introduce an even more lightweight DPO-Convex algorithm that operates on one GPU, by leveraging the convex optimization reformulation of Neural Networks (NNs). Our aim to provide faster convergence to solutions of better optimality, and higher interpretability of the underlying optimization landscape for generative language tasks. We use the Alternating Direction Method of Multipliers (ADMM) to solve this optimization problem, in order to increase parallelization efficiency. Our methods are implemented in JAX to lift the memory constraints across experiments. We experiment on three datasets, including one synthetically generated Hotel-Concierge dataset, to demonstrate the efficacy of our novel algorithm in a real world setting. Our method is comparable in user preference generation to DPO when tested on 18 human volunteers, despite being trained on one RTX-4090 with a smaller dataset.

Mentor: Archit Sharma, External Collaborators/Sharing Project: N/A

1 Introduction

Language models have been trained on increasingly large amounts of data to capture semantic language patterns. The paradigm is then a combination of pre-training and fine-tuning these LMs to achieve more preferable responses. The DPO Rafailov et al. (2024) paper proposes a simpler, performant, and computationally lightweight alternative to aligning LMs to given instructions and optimizing for human preferences. This model is able to infer what a user wants and convert that to a realistic answer that a user might like, thus taking huge steps towards inferring intent and demonstrating remarkable generalization. RLHF Wang et al. (2023) is an complex method which involves a 3-step training cycle between humans, the agent's understanding of the goal, and the Reinforcement Learning (RL) training procedure. However despite its effectiveness, RLHF is expensive, complicated, sensitive to numerous hyperparameters, and subject to unstable training procedure. The additional dependency on humans in the training loop is also unsatisfactory. This restricts RLHF to only high compute high resource companies, leading to centralized and closed AI power which is a detriment to progress. However, both of these methods typically still require multi-GPU settings to train on meaningful real world datasets.

In this project, we aim to provide a novel framework to fine-tune small language models. We use only one RTX-4090, and combine an additional auxiliary signal to the DPO loss through a convex reformulated NN. Since the convex model is hyperparameter free and of negligible cost to train, this addition is able to provide information on the optimization landscape without incurring extra cost in terms of speed. We then utilize the JAX Schoenholz and Cubuk (2020) framework to Just-In-Time compile our lower level functional code more efficiently. In order to assess the efficacy of our method, we create a synthetic Hotel-Concierge conversational dataset, and evaluate on 18 human volunteers via survey to better understand "preference".

The convex reformulation of NN for binary classification problems has been explored in our previous work. Additionally we note that DPO treats the policy optimization task as a binary classification problem, therefore the synergy is mutually advantageous. In order to practically solve the convex optimization problem, we apply ADMM with block coordinate descent for a fast and cheap method of arriving at more globally optimal solutions. Furthermore we implement our experiments in JAX, to optimize for computational cost and more efficient memory usage. As a result all of the experiments in this work were performed on one singe RTX-4090 GPU, with reasonable train times of less than 1 hour (typically within minutes), mostly across 2 datasets. Our objective is to ultimately provide an efficient way of aligning LMs to human preference that is more accessible in academic settings. We hope this can take a small step towards democratizing AI systems for the wide populace, as well as furthering learning in this exiting area.

Our main contribution is the novel DPO-Convex algorithm using JAX which trains on one GPU, for learning purposes. We also provide a custom Hotel-Concierge dataset for experimentation, and conduct human evaluation with 18 volunteers. Our key desiderata can be summarized as follows:

- Build a robust DPO algorithm that trains small LMs on one GPU with good results.
- Incorporate the Convex-NN to achieve this. Thus improving robustness, and faster convergence to solutions of better quality, with easier parallelization.
- Solve this problem with ADMM and implement our methods in JAX. This eliminates the need for complex FSDP, and also provides the advantage of more efficiently managing memory, lifting the previous large data constraints from the Convex-NN setting.

2 Related Work

Building LMs which better align to human preferences can be viewed has falling within 3 strategies. Initial algorithms of zero-shot and few-shot in-context learning Xian et al. (2017) relies on prompt engineering. Although this is able to improve the performance of LMs to produce more desired output and does not require fine-tuning, it is not able to tackle complex tasks. More sophisticated learning methods use RL to align model output with user preference. The most successful classes such as RLHF and RLAIF have been able to create conversational LM such as ChatGPT, but despite their impressive performance is extremely complex, requires humans in the loop, and incurs significant computational resources. Therefore the authors of DPO developed a simple yet extremely performant learning algorithm to directly optimize to human preferences, without explicit reward modeling. The official implementation of DPO references four 80GB A100s, which reduces the barrier to training LMs. However the performant multi-GPU cluster setting may still not be readily available to all researchers in academia, and introduces the hyperparameter beta, with two policy and reference models.

Bengio et al. (2005) have previously shown that it is possible to characterize the optimization problem for neural networks as a convex program. Pilanci and Ergen (2020) further developed exact convex reformulations of training a two-layer ReLU neural network. The basis of this representation arises from semi-infinite duality theory, and was derived by **?** using duality theory to show that two-layer neural networks with ReLU activations and weight decay regularization may be re-expressed as a linear model with a group one penalty and polyhedral cone constraints. This is a significant step towards achieving globally optimal interpretable results. The advantage of training NN with convex optimization techniques means there exists the possibility of arriving at globally optimal results with higher transparency. This yields both practical benefits in implementation and in optimizing the non-convex landscape of NN. However this framework is typically feasible on two-layer NN, and on small data such as downsampled CIAR-10 or MNIST. In order to apply this method to area of LMs where large data is paramount, we seek better solutions.

To practically solve this convex optimization problem, Bai et al. (2018) have proposed varying approaches based on the Alternating Direction Method of Multipliers (ADMM) Boyd et al. (2011). ADMM offers several attractive advantages, such as its robustness against hyperparameter selection, linear decomposability for distributed optimization, and immunity to vanishing/exploding gradients. The successful application of ADMM in solving optimization problems across a wide range of domains has been well studied. This includes diverse fields such as control theory, maximum a posteriori (MAP) inference problems, computational biology and finance. The natural parallelization aspects of ADMM seem to make it particularly suitable to deep learning problems. Therefore we aim to integrate the convex reformulation for binary classification problems framework with DPO. Our goal to have the convex guy provide more signal to the DPO loss, thus leveraging the faster convergence to solutions of better quality from convex into DPO, then solve this large data problem with ADMM.

3 DPO and Convex Neural Network Approach

Convex-NN for Convergence The seminal work of Pilanci and Ergen (2020) introduces convex duality theory for non-convex neural network objectives. These results offer a characterization of NN models by using convex regularization in a higher dimensional space, where the data matrix is divided according to all possible hyperplane arrangements. Therefore we can define the equivalent convex reformulation of a 2-layer MLP as

$$f_{\text{CVX-MLP}}(x) = \sum_{i=1}^{P_s} D_i x^T (u_j - v_j) \tag{1}$$

Where $D_i = \textbf{diag}(\mathbb{I}[X^T h_i \geq 0])$ are diagonal matrices used to sample P_s activation patterns of the ReLU network. We aim to leverage the rich representation power of this network via their universal approximation property.

The standard non-convex formulation of a two-layer ReLU-MLP with weights $w_j^{(1)}$, $w_j^{(2)}$ is:

$$f_{\text{NCVX-MLP}}(x) = \sum_{j=1}^{m} \left(x^T w_j^{(1)} \right)_+ w_j^{(2)} \tag{2}$$

$X \in \mathbb{R}^{n \times d}$ is a two-layer ReLU-MLP, usually trained in a stochastic setting. Scalable in large datasets, but sensitive to hyperparameter tuning and lacks optimality guarantees.

From a high level, Lagrangian zero duality gap perspective proves that there does exist a convex program which achieves the same optimal value as the non-convex problem. Therefore by leveraging this perspective, we reach better interpretability and understanding of the optimization landscape neural networks. Pilanci and Ergen (2020) provide the theoretical proof and analysis of the convex reformulation. So why do we desire to solve a convex problem instead? Well it is more efficient and leads to solutions that generalize well. By applying the mechanics of convex optimization to DPO, we aim to achieve better robustness, interpretability, and faster convergence to a globally optimal solution.

ADMM for Parallelism The Alternating Direction Method of Multipliers (ADMM) solves convex optimization problems by breaking them into smaller subproblems, each of which are then easier to handle. The seminal work of Boyd et al has demonstrated the versatility of this algorithm on a wide application of problems.

The recent work of Bai et al applied ADMM to solving this problem by introducing slack variables to arrive at a convex reformulation with mean squared error loss:

$$\min_{v,s,u} ||Fu - y||_2^2 + \beta ||v||_{2,1} + \mathbb{I}_{\geq 0}(s) \quad \text{s.t.} \;\; u = v, \; Gu = s \tag{3}$$

Matrix $F \in \mathbb{R}^{n \times 2dP_s}$ in this formulation is block-wise constructed by $D_i X$ terms. This optimization problem yields the (simplified) ADMM updates explicitly:

$$\textbf{Primal u update}: \quad Au^{k+1} = b \quad \text{for } A = I + \frac{1}{\rho} F^T F + G^T G \tag{4a}$$

$$\textbf{Primal v update}: \quad v^{k+1} = \textbf{prox}_{\frac{\beta}{\rho} ||\cdot||_2} (u^{k+1} + \lambda^k) \tag{4b}$$

$$\textbf{Dual update}: \quad \lambda^{k+1} = \lambda^k + \gamma_\alpha (u^{k+1} - v^{k+1}) \tag{4c}$$

In order to avoid the expensive Cholesky decomposition to solve the primal update step, we apply ADMM to take advantage its potential for acceleration and scalability. The main attraction of ADMM in this work is its decomposition into subproblems which can be solved independently and in parallel. This is particularly advantageous for large-scale distributed or high-dimensional problems, and is particularly relevant since the the importance of GPUs in language modeling cannot be overstated.

Additionally, ADMM offers satisfying analysis of convergence guarantees under mild assumptions, which is particularly desirable when working with the non-convex landscape of language models. This method is extremely robust to tuning of hyperparameters, while offering a more transparent understanding of the underlying landscape of optimization.

JAX for Speed and Memory JAX is a lover level numerical framework that offers speed and memory advantages. Since JAX was developed for high-performance machine learning research, our past work has found it to be extremely performant in GPU acceleration settings.

DPO has brought down the barrier to entry significantly for alignment and instruction fine-tuning compared to RLHF. However, naively implementing the DPO training pipeline is still computationally significant. For reference, the DPO official implementation by the authors at dpo, still requires a cluster of A100 GPUs to overcome the memory bottlenecks in deep learning. Therefore in order to implement this work within the compute resources of one GPU (section 4.2), we take advantage of JAX's JIT feature and XLA (Accelerated Linear Algebra) compiler. Recent research in review will provide more in depth discussion on the lower level optimizations of JAX. Parallelism is an additional attractive feature, since JAX supports easy parallelism over multiple devices, such as multiple GPUs or TPUs, using pmap without the need for FSDP functions.

4 Experiments

Our goal is to examine the effectiveness of DPO to train a small language model on one GPU, and to see if we can make the process even more cost effective by providing more signal with the ADMM optimized convex neural network.

4.1 Data

This study explores three datasets: both synthetically generated and well-established datasets to be consistent with previous work. Each dataset is selected to offer a different qualitative assessment of the methodology. We format each dataset into "prompt", "chosen", "rejected" labels to be consistent with the original DPO paper. Appendix A contains examples of the training dataset, as well as generated samples. In each case we follow the DPO dataset of preferences format with $\mathcal{D}$ be defined as follows: $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_i^N$. Where $y_w^{(i)}$ is the "chosen" output and $y_l^{(i)}$ is the "rejected" output.

- **IMDb Sentiment Generation** This dataset contains a collection of positive and negative movie reviews from IMDb for the task of controlled sentiment generation. This is selected as the baseline dataset for all methods, to be consistent with the original DPO paper and verify mode implementations. In this case x is the title of a movie, and y is the generated positive sentiment, which should also accurately reflect the movie.
- **Hotel Concierge Dataset** This is a custom generated task-orientated dataset in a hospitality setting. Please see Appendix A for data samples. In each conversation we create 4000 dialogue prompts with GPT4 then use 2 instances of ChatGPT to simulate conversations a hotel guest might have with a hotel concierge. The dataset is formatted as Prompt, Agent 1, Agent 2, Agent 1, etc. We then create the DPO dataset with $y_w^{(i)}$ as the completion immediately following the guests query, and $y_l^{(i)}$ of the alternative agent's generation 2 steps forwards from the guests query. The creation of this dataset serves 2 purposes: Since real world applications often provide limited or unlabeled data, we are interested in how well human preferences can be optimized with a simulated real world dataset in a well-defined hospitality setting. Secondly, since this is the smallest of our three datasets, we are interested in the possibility of aligning LM to human preference with very little data as described by the authors of LIMA. This dataset is used to generate responses to hotel guest queries/conversations as a task.

- **Stanford-SHP** This is the largest dataset in our experiments, and is selected to stress test the memory and speed performance of our models on the setup described in section 4.2. The Stanford-SHP is a dataset of 385K collective human preferences over responses to questions in 18 different subject areas. This dataset also serves to generate preferable responses to prompts, however due to the slow iteration and sample during eval limits, we are more interested in how it affects our systems compute and qualitative generative output performance.

4.2 Experimental Details

Throughout all experiments we use DistilGPT2Li et al. (2021) architecture as both the reference and policy model. Our selection is due to its versatility to run in both JAX and Pytorch frameworks, while utilizing a small number of 82 million parameters. This architecture retains approximately 97% of GPT2's language understanding skills despite its reduced size. Since our analysis is interested in how our implementations compare relative to each other, and due to limited computational resources, with aims of fast iteration this architecture was the lightest but still versatile choice. All experiments are run singularly on Ubuntu 22.04 with one RTX-4090, CUDA 12.4 and Jax 0.4.28. Maximum training time reached 2.15 hours on the Stanford-SHP with the DPO loss, while minimum training time occurred with the custom Hotel Concierge dataset in supervised fine-tuning mode of approximately 2min. We keep the same learning rate and configurations as the official DPO implementation.

4.3 DPO with Convex-NN Feedback

In this section we describe the 3 model implementations of this work. For each model, we train and evaluate on the Hotel Concierge Dataset, then the IMDb Sentiment dataset as described in section 4.1. The Stanford-SHP dataset is only trained and evaluated on the JAX-DPO re-implementation of the official source code by [dpo. All other code has been custom coded by the author of this project. During evaluation, metrics and training loss are monitored on Weights and Biases, then during human evaluation we sample from our frozen and custom trained models.

Baseline The baseline model is simple DistilGPT2 with supervised fine-tuning loss. In subsequent DPO settings, we initiate from the saved model checkpoint (policy.pt) of the SFT baseline.

DPO Model Next we train and evaluate on the DPO model. The reference model is essentially frozen, and we optimize the policy model with the DPO loss defined as follows:

$$L_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\frac{\log\pi_\theta(y_w|x)}{\log\pi_{\text{ref}}(y_w|x)} - \beta\frac{\log\pi_\theta(y_l|x)}{\log\pi_{\text{ref}}(y_l|x)}\right)\right]. \tag{5}$$

dpo provide in-depth analysis on the mechanics of DPO. This step is significantly more memory and time intensive than the baseline model. Naive implementations of the DPO model in Pytorch are not able to complete training due to compute limitations of the experiment setting in section 4.2. Therefore we reimplement this trainloop in JAX, remove FSDP, and rewrite utils to load our custom features.

DPO-Convex Model Our novel algorithm builds on prior work, where we have seen that the combination of the convex ReLU NN implemented with ADMM in JAX is able to handle datasizes such as ImageNet and IMDb and yields faster convergence with solutions of better quality. We are motivated by the DPO objective, which treats the policy optimization task as a binary classification with cross-entropy problem. Therefore, what if we can speed up the optimization of the DPO loss by giving it an additional auxiliary signal with the convex NN model?

In observing the DPO loss, we note that the main component is the inner log ratio between the policy model and the reference model. We conjecture that by extracting the hidden features as the policy model optimizes, we should be able to leverage the convex-admm method to solve the binary classification problem. The output of the convex model provides optimal weights and classification metrics, therefore we label all "chosen" = 1 and all "rejected" = 0 in training to optimize for user preferences. This convex block is then added into the training loop of the DPO training pipeline, and used to optimize DPO's BCE style loss.

The official implementation of DPO uses RMSProp, which is seen to be as performant as Adam but more memory efficient since it requires less storage variables. However we note that the integration of the convex-DPO algorithm can provide advantages such as robustness against hyperparameter tuning and faster convergence. This aims to push the DPO loss towards a more globally optimal solution even faster. Please see appendix B for performance plots.

4.4 Evaluation method

Evaluation of generative language models is an activate area of research and an extremely difficult task. This is because human preference is hard to define, and recent work of Archit has shown that often humans will prefer simply the longer generated output without reason. Therefore we qualitatively evaluate our output with the 17 human participants. This is also in order to be consistent with existing literature of the seminal DPO paper. We structure evaluation as follows:

- vary temperature hyperparameter from 0.001 to 2.101 in steps of 0.3.
- for each temperature, in each of the 3 models listed in section 4.3 above, we input the same 7 prompts. For example, a prompt might be "Where is the nearest park?"
- each of the models generate a response, which is shuffled into a multiple choice survey, and sent to 17 human volunteers
- This is repeated for the IMDb dataset with the task of positive sentiment completion

Detailed results and samples of survey output, as well as acknowledgements to participants are listed in Appendix C. Table 4.5 summarizes the average performance of each model. The IMDb prompt survey instructed users to rate the movie review they found the most POSITIVE. The description of this setting is "You are browsing for a new movie to watch". The Hotel Concierge response survey section asked users to select the response that was the most HELPFUL and HUMAN. this section was described to users as "you are checking into a hotel and conversing with the concierge".

We also measure the speed and stability of training, as well as the robustness to hyperparameter tuning on the convex setting.

Finally we observe training time, loss achieved (see below), and difference in scalability between frameworks.

4.5 Results

In this section we compare the results of the 3 models discussed in section 4.3. Although we perform ablation studies with varying $0.001 < T < 2.01$. The baseline model is consistently the fastest to train, although it consistently demonstrates the highest amount of repetition in its output. This is further validated in our human feedback survey, where the baseline model won on only one out of thirteen questions.

The DPO-Convex model shows the most stable training performance. Despite variances in hyperparameters such as temperature, data size, batch size, this model was consistently able to stably and quickly decrease in loss. Figures 1 and Figure 2 show the training performance of the DPO-Convex model without any tuning of hyperparameters. On the same dataset, the DPO-Convex model is significantly faster to train than the DPO model under the same conditions. For example, the naive DPO model itself when trained on the Hotel-Concierge dataset needed approximately 1 hour, however the DPO-Convex model was able to complete training in significantly less time (30min). We attribute this to the more efficient implementation of of the DPO-Convex model in JAX, and its efficacy at solving the Convex-NN problem. The stable and fast training aspects of the DPO-Convex model is attractive, and leaves room for further experimentation.

The generative ability and ratio of preference win rate between the DPO and DPO-Convex model are almost equivalent in our human feedback survey. We note that the DPO model is the most sensitive of the 3 methods to varying T, and we believe that larger datasets with longer training and more epochs will likely yield significant differences between these two models. Table 4.5 summarizes the results of the human feedback survey, and shows both win rate and the average preference of each model. The average preference is calculated as the percentage of each model's win rate divided by the number of times it won. We provide the average preference percentage as a metric since it gives

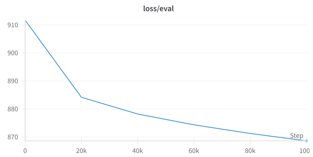


Figure 1: DPO-Convex Evaln

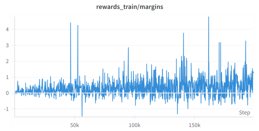


Figure 2: DPO-Convex Train

Table 1: Feedback from 18 Human Volunteers

	FST	DPO	DPO-C
IMDb (Win rate)	1	3	4
Hotel (Win rate)	0	3	3
IMDb Avg Win %	72.2%	62.5%	68.5
Hotel Avg Win %	0	68.7%	83.3%

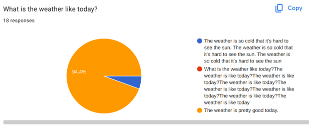
better signal as to how preferred a model was. For example, the baseline FST model only won on one question, but was strongly preferred in that case by most humans. The Hotel-Concierge dataset saw an equal win-rate count between the DPO model and DPO-Convex model, but humans had stronger preference to the answers of DPO-Convex (83.3%).

5 Analysis and Discussion

Since we use smaller datasets on the DistilGPT2Li et al. (2021) model, we expected to see a certain amount of repetition in the output. This is most prominent in the baseline FST model. For example, the prompt "How can I check in?The answer is yes. I can't....". Although we vary Temperature and its effect on perplexity from 0.001 to 2.01 in steps of 0.3, the baseline FST does not increase in performance and is consistent in its repetition (as seen in B).

The DPO model needed approximately double the amount of time to train as the baseline FST model. However the DPO model notably generated varying degrees of creativity in the same prompt as temperature varied. We note that the DPO model instances that won on the human feedback survey were all instances where $T < 1$. This is in contrast to our conjecture that higher temperature will produce more desirable results with DPO since humans prefer more creative output. We also note that since our generated dataset is in a Hotel-Concierge setting, it's possible humans prefer more consistency versus creativity. In the two sample questions posed to our human volunteers, it is clearly seen that the baseline model shows repetition, but the DPO model is preferred with $T = 0.601$. The DPO-Convex model tends to generate longer responses. However this might be attributed to its capacity for faster training.

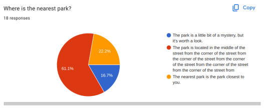
The DPO-Convex model showed the most stable training performance. While training on one GPU and without compromising dataset size, loss was able to consistently go down regardless of varying hyperparameters. This agrees with our conjecture that adding the convex feedback increases robustness, and eliminated the need to continue with further hyperparameter tuning in experiments with the DPO-convex model. Please see Appendix B for training plots. In human feedback, both DPO and the DPO-Convex model were almost equally preferred. We attribute this to small sample size of questions and volunteers, and realize the significance and difficulty of evaluating preference generation. This direction leaves room for more future work.


What is the weather like today?

18 responses

- ● The weather is so cold that it's hard to see the sun. The weather is so cold that it's hard to see the sun. The weather is so cold that it's hard to see the sun
- ● What is the weather like today?The weather is like today?The weather is like today?The weather is like today?The weather is like today?The weather is like today?The weather is like today?The weather is like today
- ● The weather is pretty good today.

Figure 3: DPO wins with T=0.601


Where is the nearest park?

18 responses

- ● The park is a little bit of a mystery, but it's worth a look.
- ● The park is located in the middle of the street from the corner of the street from the corner of the street from the corner of the street from the corner of the street from the corner of the street from
- ● The nearest park is the park closest to you.

Figure 4: DPO-Convex wins

6 Conclusion

We have shown that it is possible to provide extra signal to the DPO loss by leveraging the convex reformulation of a two layer neural network. This novel algorithm seeks to combine the robustness and faster convergence of the convex auxiliary signal with the DPO objective. The resulting algorithm is more robust to hyperparameter tuning (such as temperature), and allows quick iteration with preferable output. We implement our methods in JAX such that all experiments run on one GPU for speed and better memory efficiency, and provide a synthetic dataset of a Hotel-Concierge setting for analysis. Thus we hope this work can reduce the barrier for entry even more for individual researchers, in the exciting field of optimizing LMs for human preferences.

Limitations and Future Work Future work will involve running our JAX experiments on TPUs or GPU clusters. Since JAX was developed with easy parallelization in mind, more performant scaling results should be explored with we can handle even more data. Better analysis of the theoretical implications of the DPO-Convex algorithm is desired, and other options of optimizing convex NN problems should be explored.

7 Ethics Statement

Bias in responses of generated language is a major ethical challenge. In learning from public datasets on hotels or sentiment, the model may inadvertently develop negative biases which can lead to discriminatory behavior. Examples of this include: favoring or disfavoring certain groups of users over others, showing bias in service recommendations based on perceived socioeconomic background, or pushing certain services based on biased signals in Name origins. Secondly, any application in sentiment analysis has the raising concern of psychological manipulation and user autonomy. Since users often don't know what's most helpful to themselves, it's possible for the model to induce unhealthy psychological effects without knowledge of consequences.

To mitigate these risks, we should implement strategies to ensure fairness and curate training data to represent a diverse range of user interactions from various demographic groups. Furthermore users should be given notice prior to engaging with the model that all generated feedback should be viewed as ML feedback, and give room for independent thought. Clear opt-in consent for users should be presented at the onset. Finally regular audits should be conducted to check for any biases that arise, and feedback from users should be regularly collected to ensure ongoing fairness.

8 Acknowledgements

We thank Professor Manning, our project TA Archit Sharma, Anna Goldie, Kevin Nam, and the entire 224N teaching team! We also gratefully acknowledge Lucy, who has stayed up with us on many late nights working on this project.

References

Dpo: Direct preference optimization. In *https://github.com/eric-mitchell/direct-preference-optimization/tree/main*.

Anthropic. https://www.anthropic.com/claude.

Jianchao Bai, Jicheng Li, Fengmin Xu, and Hongchao Zhang. 2018. Generalized symmetric admm for separable convex optimization. *Computational optimization and applications*, 70(1):129–170.

Yoshua Bengio, Nicolas Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte. 2005. Convex neural networks. *Advances in neural information processing systems*, 18.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

Tianda Li, Yassir El Mesbahi, Ivan Kobyzev, Ahmad Rashid, Atif Mahmud, Nithin Anchuri, Habib Hajimolahoseini, Yang Liu, and Mehdi Rezagholizadeh. 2021. A short study on compressing decoder-based language models. *arXiv preprint arXiv:2110.08460*.

Inc. OpenAI. https://openai.com/chatgpt/.

Mert Pilanci and Tolga Ergen. 2020. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *International Conference on Machine Learning*, pages 7695–7705. PMLR.

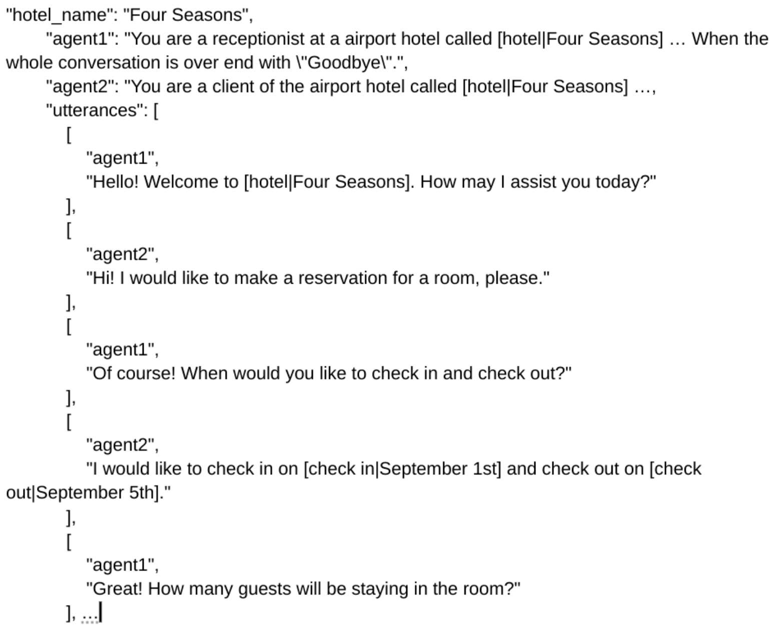
Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Samuel Schoenholz and Ekin Dogus Cubuk. 2020. Jax md: a framework for differentiable physics. *Advances in Neural Information Processing Systems*, 33:11428–11441.

Yuanhao Wang, Qinghua Liu, and Chi Jin. 2023. Is rlhf more difficult than standard rl? *arXiv preprint arXiv:2306.14111*.

Yongqin Xian, Bernt Schiele, and Zeynep Akata. 2017. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4582–4591.

A Examples of Data



```
"hotel_name": "Four Seasons",
        "agent1": "You are a receptionist at a airport hotel called [hotel|Four Seasons] … When the
whole conversation is over end with \"Goodbye\".",
        "agent2": "You are a client of the airport hotel called [hotel|Four Seasons] …,
        "utterances": [
            [
                "agent1",
                "Hello! Welcome to [hotel|Four Seasons]. How may I assist you today?"
            ],
            [
                "agent2",
                "Hi! I would like to make a reservation for a room, please."
            ],
            [
                "agent1",
                "Of course! When would you like to check in and check out?"
            ],
            [
                "agent2",
                "I would like to check in on [check in|September 1st] and check out on [check
out|September 5th]."
            ],
            [
                "agent1",
                "Great! How many guests will be staying in the room?"
            ], ...
```

Figure 5: Example of custom generated Hotel Concierge dataset

B Performance Plots

Please see the following images for performance plots.

C Human Evaluation

We have 18 human volunteers selecting their most preferred generated output. The survey is conducted as a total of 13 questions, across 2 datasets (4.1), with output generated by each of the three models discussed in 4.3. Further details of survey human evlauation is summarized in 4.5. The raters were Stanford students (from graduate to Ph.D.), University of Toronto students (Ph.D.), Google

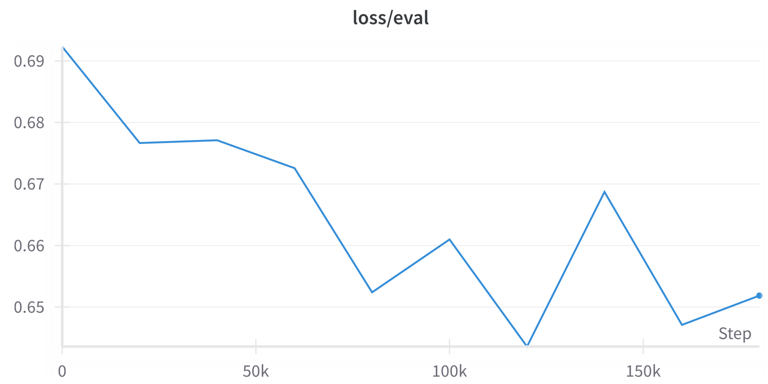


Figure 6: Training for DPO naive

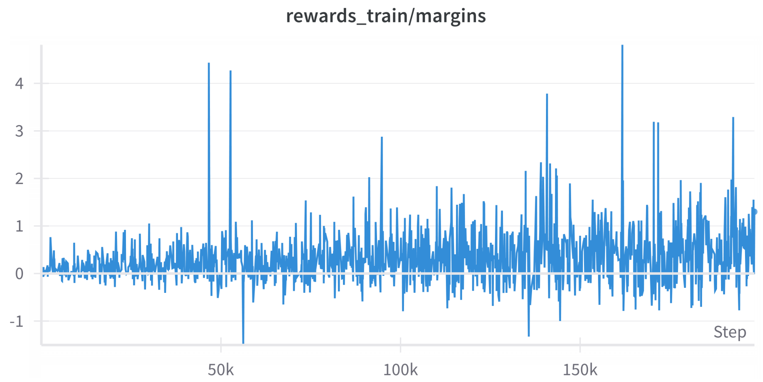


Figure 7: Larger image of training for DPO-Convex as discussed above

Software Engineers, and medical practitioners in Veterinary science. We gratefully acknowledge the contribution of each of our volunteers, listed in random order: Anna Goldie, Kevin Nam, Zhong Wei Dang, Shaun Benjamin, Sera Benjamin, Yue Benjamin, Noriyuki Shintoku, Jenny Song, Zachary Frangella, Stephen Sapperton, Mary Habib, Scarlet Arreola Barrones, Farhis Kordi, Adam Wong, Jia Yu, Tong Zhang, Juan Jaramillo, and Trevor Germain.

D Results

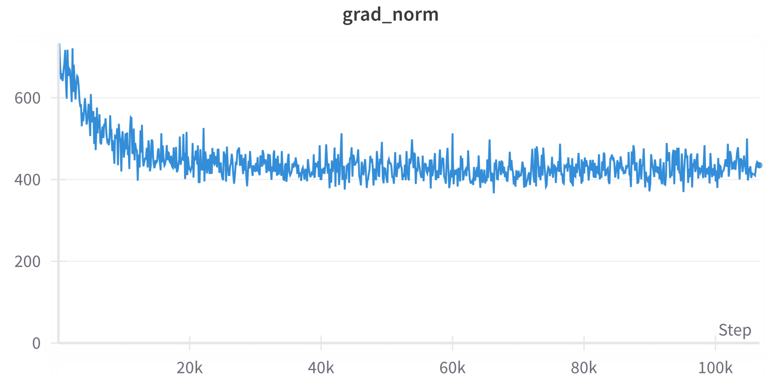
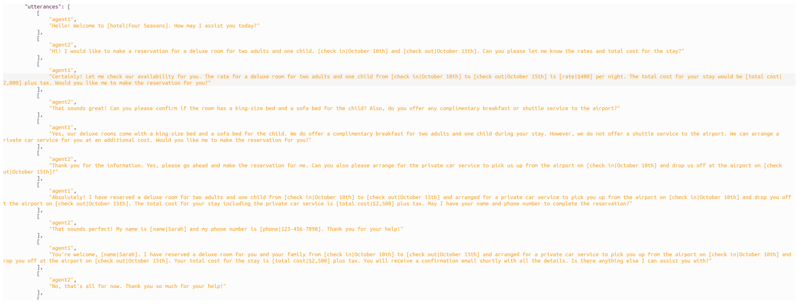
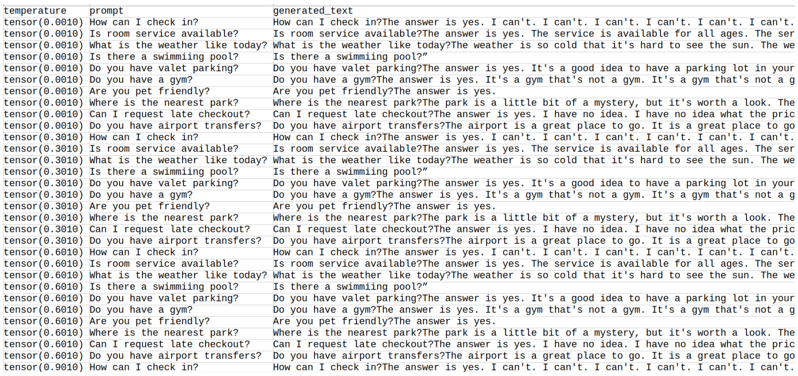


Figure 8: Sample of training for DPO-Convex



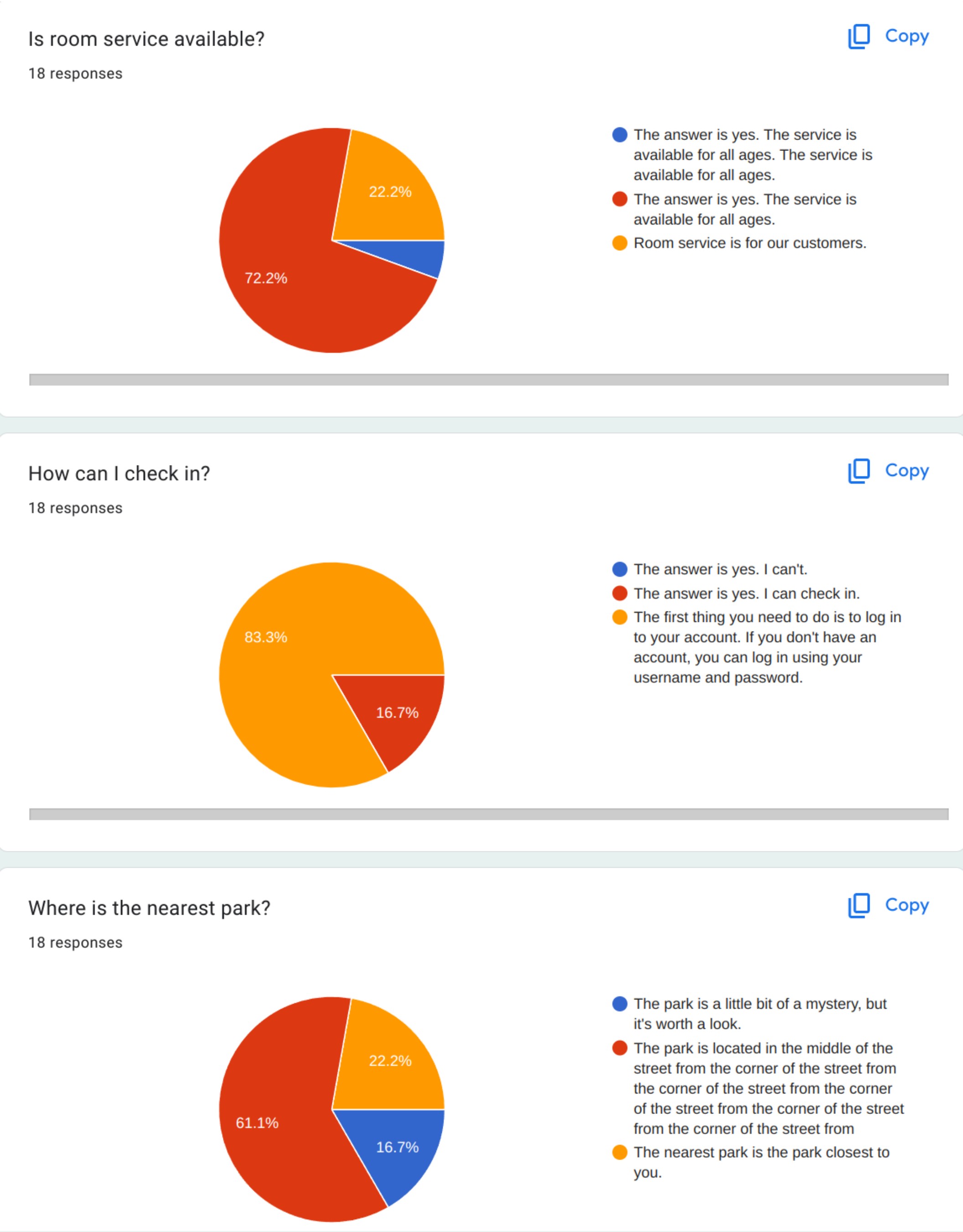
```
    "utterances": [
        [
            "agent1",
            "Hello! Welcome to [hotel|Four Seasons]. How may I assist you today?"
        ],
        [
            "agent2",
            "Hi! I would like to make a reservation for a deluxe room for two adults and one child. [check in|October 10th] and [check out|October 15th]. Can you please let me know the rates and total cost for the stay?"
        ],
        [
            "agent1",
            "Certainly! Let me check our availability for you. The rate for a deluxe room for two adults and one child from [check in|October 10th] to [check out|October 15th] is [rate|$400] per night. The total cost for your stay would be [total cost|
2,000] plus tax. Would you like me to make the reservation for you?"
        ],
        [
            "agent2",
            "That sounds great! Can you please confirm if the room has a king-size bed and a sofa bed for the child? Also, do you offer any complimentary breakfast or shuttle service to the airport?"
        ],
        [
            "agent1",
            "Yes, our deluxe rooms come with a king-size bed and a sofa bed for the child. We do offer a complimentary breakfast for two adults and one child during your stay. However, we do not offer a shuttle service to the airport. We can arrange a
rivate car service for you at an additional cost. Would you like me to make the reservation for you?"
        ],
        [
            "agent2",
            "Thank you for the information. Yes, please go ahead and make the reservation for me. Can you also please arrange for the private car service to pick us up from the airport on [check in|October 10th] and drop us off at the airport on [check
ut|October 15th]?"
        ],
        [
            "agent1",
            "Absolutely! I have reserved a deluxe room for two adults and one child from [check in|October 10th] to [check out|October 15th] and arranged for a private car service to pick you up from the airport on [check in|October 10th] and drop you off
t the airport on [check out|October 15th]. The total cost for your stay including the private car service is [total cost|$2,500] plus tax. May I have your name and phone number to complete the reservation?"
        ],
        [
            "agent2",
            "That sounds perfect! My name is [name|Sarah] and my phone number is [phone|123-456-7890]. Thank you for your help!"
        ],
        [
            "agent1",
            "You're welcome, [name|Sarah]. I have reserved a deluxe room for you and your family from [check in|October 10th] to [check out|October 15th] and arranged for a private car service to pick you up from the airport on [check in|October 10th] and
rop you off at the airport on [check out|October 15th]. Your total cost for the stay is [total cost|$2,500] plus tax. You will receive a confirmation email shortly with all the details. Is there anything else I can assist you with?"
        ],
        [
            "agent2",
            "No, that's all for now. Thank you so much for your help!"
        ],
        [
```

Figure 9: Sample of Survey to Human Volunteers



temperature	prompt	generated_text
tensor(0.0010)	How can I check in?	How can I check in?The answer is yes. I can't. I can't. I can't. I can't. I can't. I can't.
tensor(0.0010)	Is room service available?	Is room service available?The answer is yes. The service is available for all ages. The ser
tensor(0.0010)	What is the weather like today?	What is the weather like today?The weather is so cold that it's hard to see the sun. The we
tensor(0.0010)	Is there a swimmiing pool?	Is there a swimmiing pool?”
tensor(0.0010)	Do you have valet parking?	Do you have valet parking?The answer is yes. It's a good idea to have a parking lot in your
tensor(0.0010)	Do you have a gym?	Do you have a gym?The answer is yes. It's a gym that's not a gym. It's a gym that's not a g
tensor(0.0010)	Are you pet friendly?	Are you pet friendly?The answer is yes.
tensor(0.0010)	Where is the nearest park?	Where is the nearest park?The park is a little bit of a mystery, but it's worth a look. The
tensor(0.0010)	Can I request late checkout?	Can I request late checkout?The answer is yes. I have no idea. I have no idea what the pric
tensor(0.0010)	Do you have airport transfers?	Do you have airport transfers?The airport is a great place to go. It is a great place to go
tensor(0.3010)	How can I check in?	How can I check in?The answer is yes. I can't. I can't. I can't. I can't. I can't. I can't.
tensor(0.3010)	Is room service available?	Is room service available?The answer is yes. The service is available for all ages. The ser
tensor(0.3010)	What is the weather like today?	What is the weather like today?The weather is so cold that it's hard to see the sun. The we
tensor(0.3010)	Is there a swimmiing pool?	Is there a swimmiing pool?”
tensor(0.3010)	Do you have valet parking?	Do you have valet parking?The answer is yes. It's a good idea to have a parking lot in your
tensor(0.3010)	Do you have a gym?	Do you have a gym?The answer is yes. It's a gym that's not a gym. It's a gym that's not a g
tensor(0.3010)	Are you pet friendly?	Are you pet friendly?The answer is yes.
tensor(0.3010)	Where is the nearest park?	Where is the nearest park?The park is a little bit of a mystery, but it's worth a look. The
tensor(0.3010)	Can I request late checkout?	Can I request late checkout?The answer is yes. I have no idea. I have no idea what the pric
tensor(0.3010)	Do you have airport transfers?	Do you have airport transfers?The airport is a great place to go. It is a great place to go
tensor(0.6010)	How can I check in?	How can I check in?The answer is yes. I can't. I can't. I can't. I can't. I can't. I can't.
tensor(0.6010)	Is room service available?	Is room service available?The answer is yes. The service is available for all ages. The ser
tensor(0.6010)	What is the weather like today?	What is the weather like today?The weather is so cold that it's hard to see the sun. The we
tensor(0.6010)	Is there a swimmiing pool?	Is there a swimmiing pool?”
tensor(0.6010)	Do you have valet parking?	Do you have valet parking?The answer is yes. It's a good idea to have a parking lot in your
tensor(0.6010)	Do you have a gym?	Do you have a gym?The answer is yes. It's a gym that's not a gym. It's a gym that's not a g
tensor(0.6010)	Are you pet friendly?	Are you pet friendly?The answer is yes.
tensor(0.6010)	Where is the nearest park?	Where is the nearest park?The park is a little bit of a mystery, but it's worth a look. The
tensor(0.6010)	Can I request late checkout?	Can I request late checkout?The answer is yes. I have no idea. I have no idea what the pric
tensor(0.6010)	Do you have airport transfers?	Do you have airport transfers?The airport is a great place to go. It is a great place to go
tensor(0.9010)	How can I check in?	How can I check in?The answer is yes. I can't. I can't. I can't. I can't. I can't. I can't.

Figure 10: FST baseline resistant to varying T, consistently repeats output



Is room service available?

18 responses

● The answer is yes. The service is available for all ages. The service is available for all ages.

● The answer is yes. The service is available for all ages.

● Room service is for our customers.

How can I check in?

18 responses

● The answer is yes. I can't.

● The answer is yes. I can check in.

● The first thing you need to do is to log in to your account. If you don't have an account, you can log in using your username and password.

Where is the nearest park?

18 responses

● The park is a little bit of a mystery, but it's worth a look.

● The park is located in the middle of the street from the corner of the street from the corner of the street from the corner of the street from the corner of the street from the corner of the street from

● The nearest park is the park closest to you.

Figure 11: Sample of survey output from human feedback