P3O: Pessimistic Preference-based Policy Optimization for Robust Alignment from Preferences

Dhawal Gupta[*]
University of Massachusetts
dgupta@cs.umass.edu

Christoph Dann
Google Research
chrisdann@google.com

Alekh Agarwal
Google Research
alekhagarwal@google.com

Abstract

We study reinforcement learning (RL) settings where the agent only has access to preferences on the relative quality of a pair of trajectories, obtained as a fixed *offline preference dataset*, where pairs of trajectories collected according to some base policy are labeled with the preference feedback. A reward or pairwise preference function trained from this offline dataset is then used to provide feedback during RL training, and there is a substantial body of work on RL methods for these settings. However, a bulk of the literature ignores the uncertainty of the learned preference function, which leads to reward hacking or overoptimization. In this work, we formulate theoretically sound objectives for preference-based RL (PbRL) which are provably robust to overoptimization through the use of pessimism in the face of uncertainty, and design practical algorithms to optimize these objectives. We evaluate our algorithms on the task of fine-tuning language models from human feedback, and show a remarkable resilience to overoptimization.

1 Introduction

Reinforcement learning from human feedback (RLHF) [Christiano et al., 2017] has emerged as a promising technique for aligning language models with human preferences [Stiennon et al., 2020, Ouyang et al., 2022]. The predominant approach involves training a reward model on human preference data and then fine-tuning the language model to maximize this reward. More recently, a line of works [Swamy et al., 2024, Munos et al., 2023, Calandriello et al., 2024, Guo et al., 2024] argue for the benefits of learning a pairwise preference function from the preference dataset, and using this to compare trajectories during online RL. Irrespective of whether we use reward or preference models during subsequent RL, however, the availability of a limited pool of high-quality preference dataset presents a key bottleneck in learning good policies. The high cost of collecting preference datasets means that they suffer from limited coverage, and models trained on such datasets fail to adequately generalize to policies which produce trajectories out of the support of the preference data.

The inadequacy of learned reward/preference models in reliably producing good policies has resulted in the now well-documented phenomenon of reward hacking or overoptimization [Amodei et al., 2016, Gao et al., 2023, Eisenstein et al., 2024]. Correspondingly, there is a growing literature on techniques to control this overoptimization behavior, such as by incorporating uncertainty in the predictions of the underlying reward model using explicit reward ensembles [Eisenstein et al., 2024, Coste et al., 2023], or pessimistic reasoning [Fisch et al., 2024, Liu et al., 2024, Huang et al., 2024b, Cen et al., 2024]. In contrast to the reward-based setting, much less work studies the incorporation of uncertainty when using learned pairwise preference models in subsequent RL.

[*]The work was done as a student researcher at Google Research (dhawgupta@google.com).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

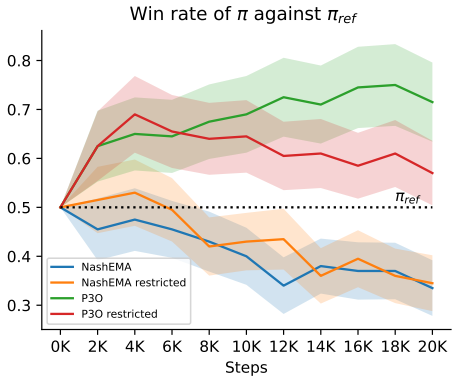

Figure 1: The horizontal axis corresponds to number of learning steps, and the vertical axis represents the preference value evaluated against $\pi_{\texttt{ref}}$ using Gemini 1.0 Ultra. Shaded areas represent $\pm 1\times$ standard error (see Sec. 4 for details).

In this work, we build on prior works in preference-based RLHF [Swamy et al., 2024, Munos et al., 2023], as well as offline learning in Markov games [Cui and Du, 2022], to obtain new robust objectives for incorporating uncertainty from finite preference datasets, and make the following contributions:

1. We point out undesirable properties of the most natural pessimistic estimator motivated by Cui and Du [2022], when the offline dataset has some systematic gaps in its coverage. We develop a new formulation under which the learned policy is provably preferable to any other policy which chooses actions in the support of the dataset, and show the theoretical benefits of this formulation.
2. We provide a practical algorithm for optimizing the resulting objective. Existing approaches for preference-based RLHF [Swamy et al., 2024, Munos et al., 2023] already involve a minimax game, so adding further minimization over preference functions in pessimism creates a challenging optimization problem. We approximate the ideal objective with a variational upper bound, that yields a minimax game between a policy and a preference player, which we solve using gradient ascent-descent. The policy optimization is similar to prior works [Swamy et al., 2024, Munos et al., 2023] and the preference updates are adversarial to the current policy's choices.
3. In a document summarization task, we find in Figure 1 that while preference-based methods without pessimism (NashEMA [Munos et al., 2023] and NashEMA restricted) exhibit significant overoptimization, our algorithms (P3O and P3O restricted) learn a good policy as evaluated by a prompted Gemini model, and their performance does not deteriorate over the training process.

2 Background

We consider human alignment of a language policy $\pi : \mathcal{X} \to \triangle(\mathcal{Y})$[2] which generates for a given context $\boldsymbol{x} \in \mathcal{X}$ a response $\boldsymbol{y} \sim \pi(\cdot|\boldsymbol{x})$ with $\boldsymbol{y} \in \mathcal{Y}$. We are given access to a preference dataset, $\mathcal{D}$, consisting of tuples $(\boldsymbol{x}, \boldsymbol{y}_{\mathrm{W}}, \boldsymbol{y}_{\mathrm{L}}) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ where for context $\boldsymbol{x}$, the response $\boldsymbol{y}_{\mathrm{W}}$ is preferred over $\boldsymbol{y}_{\mathrm{L}}$ as labeled by a human. We further assume access to a reference policy $\pi_{\texttt{ref}}$, which may or may not match the sampling policy for $\boldsymbol{y}_{\mathrm{W}}, \boldsymbol{y}_{\mathrm{L}}$. For brevity, we drop the context $\boldsymbol{x}$ from the notation and work with a finite $\mathcal{Y}$ when there is no ambiguity.

Preferences are often modeled via a reward function with the Bradley-Terry model [Christiano et al., 2017, Ouyang et al., 2022]; however, in this paper, we make no such assumptions and work with general preference functions. We first set up the preference learning framework, and then discuss techniques to optimize with preference feedback, while also establishing the use of pessimism to handle uncertainties that may exist in the reward and preference functions.

Preference Learning: We define the preference function $p : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$, such that $p(\boldsymbol{y}_1, \boldsymbol{y}_2) \doteq \Pr(\boldsymbol{y}_1 \succ \boldsymbol{y}_2)$ represents the probability of the generation $\boldsymbol{y}_1$ being preferred over $\boldsymbol{y}_2$. The preference function satisfies: $p(\boldsymbol{y}_1, \boldsymbol{y}_2) = 1 - p(\boldsymbol{y}_2, \boldsymbol{y}_1)$. To obtain a preference model, we typically fine-tune a pretrained language model (LM) on $\mathcal{D}$ to produce the maximum likelihood model p_{MLE} via the following objective:

$$p_{\mathrm{MLE}} \in \arg\min_p \mathcal{L}_{\mathrm{pref}}(p; \mathcal{D}) \;\text{ where }\; \mathcal{L}_{\mathrm{pref}}(p, \mathcal{D}) = -\tfrac{1}{|\mathcal{D}|} \textstyle\sum_{(\boldsymbol{y}_{\mathrm{W}}, \boldsymbol{y}_{\mathrm{L}}) \in \mathcal{D}} \log p\left(\boldsymbol{y}_{\mathrm{W}}, \boldsymbol{y}_{\mathrm{L}}\right). \quad (1)$$

[2]We use $\triangle(\mathcal{Y})$ to denote the probability simplex defined over the elements of the set $\mathcal{Y}$.

We overload the notation of $p(\pi, \pi')$, where $\pi, \pi' \in \triangle(\mathcal{Y})$, to represent the expected preference of π over π', given the preference function p function.

Preference Optimization: We define a regularized *maximin* [Swamy et al., 2024, Munos et al., 2023] preference game objective $J_{\mathrm{P}}(\pi, \pi', p)$ between a pair of competing policies π and π', with preference function p, a reference policy $\pi_{\texttt{ref}}$, and a regularization parameter $\tau > 0$, as

$$J_{\mathrm{P}}(\pi, \pi', p) \doteq p(\pi, \pi') - \tau \mathrm{KL}\left(\pi \| \pi_{\texttt{ref}}\right) + \tau \mathrm{KL}\left(\pi' \| \pi_{\texttt{ref}}\right),$$

where $\mathrm{KL}\left(\pi \| \pi_{\texttt{ref}}\right) \doteq \mathbb{E}_{Y \sim \pi}\left[\log \frac{\pi(Y)}{\pi_{\texttt{ref}}(Y)}\right]$. Alternatively, in the reward setting, given a reward function $r : \mathcal{Y} \to \mathbb{R}$, the objective $J_{\mathrm{R}}(\pi, r)$ is defined as: $J_{\mathrm{R}}(\pi, r) \doteq \mathbb{E}_{Y \sim \pi}[r(Y)] - \tau \mathrm{KL}\left(\pi \| \pi_{\texttt{ref}}\right)$. For the preference objective J_{P}, the π and π' players optimize their corresponding objectives, i.e.,

$$\pi_\star \in \arg\max_{\pi} \min_{\pi'} J_{\mathrm{P}}(\pi, \pi', p), \quad \pi'_\star \in \arg\min_{\pi'} \max_{\pi} J_{\mathrm{P}}(\pi, \pi', p).$$

Here, due to the symmetry of the game, a Nash equilibrium exists at the same policy, i.e., $\pi_\star = \pi'_\star$ and the objective can be simplified to a single-player game [Swamy et al., 2024], which is termed self-play preference optimization (SPO).

Pessimism in Preference Optimization: It is a well-understood issue in preference optimization and RLHF that optimizing J_{P} and J_{R} can lead to over-optimization of the corresponding preference and reward functions, resulting in a shift in the distribution of outputs [Gao et al., 2023] generated by the learned policies. These policies start exploiting regions of the preference and reward functions where uncertainty is higher, which can result in spurious high-reward areas, a phenomenon often termed "**reward hacking**". Pessimism in both the reward setting [Eisenstein et al., 2024, Liu et al., 2024, Fisch et al., 2024, Cen et al., 2024] and the preference setting [Ye et al., 2024] has been proposed as a way to remedy these issues. In this approach, we learn a policy against an adversarial reward or preference function, thus producing more robust policies.

Pessimism in the reward setting leads to a max-min game, i.e., $\pi_\star \in \arg\max_\pi \min_{r \in \mathcal{R}} J_{\mathrm{R}}(\pi, r)$ where $\mathcal{R}$ is an uncertainty set of reward functions, that is, all reward functions that are consistent with the dataset. Liu et al. [2024] and Fisch et al. [2024] show that for certain choices of $\mathcal{R}$, this game can be solved without actually maintaining the set $\mathcal{R}$ and performing the inner optimization implicitly.

In the preference optimization setting, a pessimistic solution can be naturally formulated analogously

$$\pi_\star \in \arg\max_{\pi} \min_{\pi'} \min_{p \in \mathcal{P}} J_{\mathrm{P}}(\pi, \pi', p) \tag{2}$$

where $\mathcal{P}$ defines an uncertainty set over preference functions. This formulation has been studied previously for certain choices of $\mathcal{P}$ in the tabular [Cui and Du, 2022] and function approximation setting [Ye et al., 2024, Huang et al., 2024a]. These works provide theoretical analyses showing that the solution $\pi_\star$ converges to the optimal policy as long as a condition called *unilateral coverage* holds, and show further that this condition is necessary. This approach has not been empirically evaluated in prior works, as the optimization problem is very challenging with no obvious practical strategies.

3 Method

While previous works show that (2) is a principled approach to pessimistic preference optimization with strong guarantees, this formulation has two limitations which we will address now.

3.1 Restricting the opponent to covered generations

We motivate our approach with an example which is emblematic of typical RLHF scenarios. Consider a case with no context and $\mathcal{Y} = \{y_1, y_2, y_3\}$. Suppose further that we have that $p(y_1, y_2) = 1$ for all $p \in \mathcal{P}$, so we are fully certain about this preference. But we never observe any comparisons involving y_3 in our preference data ($\pi_{\text{sample}}(y_3) = 0$), and hence the set $\mathcal{P}$ allows all values $p(y, y_3) \in [0, 1]$ for $y \neq y_3$. To highlight the limitations of pessimism in preference optimization, we consider the problem in absence of regularization, i.e., $\tau = 0$. Then, as illustrated in Figure 2 and proven in Appendix A, the optimal policy $\pi^\star$ satisfies $\pi^\star(y_3) \geq 0.5$. That is, we take an action completely out of the support of the sampled dataset w.p. ≥ 0.5, where the preferences can take

Algorithm 1: Pessimistic Preference-based Policy Optimization (P3O)

Initialize $\bar{\pi}_1 = \pi_1 = \pi_{\texttt{ref}}$ and $p_1 = p_{\text{MLE}}$;
for $t = 1, 2, \ldots$ **do**
  Set $\pi_{\text{mix}} \propto \sqrt{\bar{\pi}_t \pi_{\texttt{ref}}}$ as mix of $\pi_{\texttt{ref}}$ and EMA $\bar{\pi}_t$ for restricted Nash or $\pi_{\text{mix}} = \bar{\pi}_t$ otherwise;
  Approximate current objective

$$J(\pi_t, p_t) \doteq p_t(\pi_t, \pi_{\text{mix}}) - \tau \text{KL}\left(\pi_t \| \pi_{\texttt{ref}}\right) + \lambda \mathbb{E}_{y,y' \sim \pi_{\texttt{ref}}}\left[\text{KL}\left(p_{\text{MLE}}(y, y') \| p_t(y, y')\right)\right] \quad (3)$$

  Update: $\pi_{t+1} \leftarrow \pi_t + \eta_\pi \partial J(\pi, p_t)/\partial \pi \mid_{\pi=\pi_t}$;
  Update: $p_{t+1} \leftarrow p_t - \eta_p \partial J(\pi_t, p)/\partial p \mid_{p=p_t}$;
  Update: $\bar{\pi}_{t+1} \leftarrow \gamma \pi_t + (1-\gamma)\bar{\pi}_t$;

completely arbitrary values. In most practical applications, many possible $\boldsymbol{y}$'s will not be covered in the dataset, even distributionally, and it appears undesirable that the optimal policy obtained by pessimism predominantly generates such outputs. We now propose a remedy for this issue.

Restricted Nash for the Opponent Player

Given the example from Figure 2, an intuitive response is to consider a Nash strategy where the support for both policies is restricted to actions which are well-sampled in the preference dataset. In Appendix B, we define such a restricted Nash strategy for the function approximation case and provide theoretical guarantees for it. We also explain in Appendix C how the bounds significantly improves upon those for the unrestricted case in (2).

For a practical algorithm, we do this restriction in an approximate manner by adding an additional KL regularization term, encouraging π' to be close to the data sampling policy π_{sample}. We note that if π_{sample} is similar to $\pi_{\texttt{ref}}$, then the additional term may not be needed.

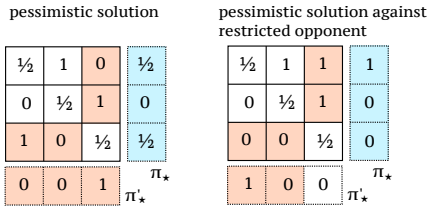

Figure 2: Problematic example for pessimism in preference optimization against unrestricted opponent. Blue and red shaded entries represent optimization variables by the max and min player respectively.

$$\max_{\pi} \min_{p \in \mathcal{P}} \min_{\pi'} p(\pi, \pi') - \tau \text{KL}\left(\pi \| \pi_{\texttt{ref}}\right) + \tau \text{KL}\left(\pi' \| \pi_{\texttt{ref}}\right) + \tau \text{KL}\left(\pi' \| \pi_{\text{sample}}\right). \quad (4)$$

Using a closed-form solution to the inner $\max_{\pi'}$ KL-regularized problem, we obtain the following lemma on an equivalent objective for π. We define the shorthand $\pi_{\text{mix}}(y; \pi_1, \pi_2) \propto \sqrt{\pi_1(y)\pi_2(y)}$, and use $y \sim \pi_{\text{mix}}(\pi_1, \pi_2)$ to abbreviate $y \sim \pi(\cdot; \pi_1, \pi_2)$.

Lemma 1. *The optimization problem* (4) *is equivalent to the following objective, assuming that the minimization over π' is over all possible policies in $\Delta(\mathcal{Y})$,*

$$\max_{\pi} \min_{p \in \mathcal{P}} - \log \mathbb{E}_{y \sim \pi_{mix}(\pi_{ref}, \pi_{sample})}\left[\exp\left(\frac{-p(\pi, y)}{2\tau}\right)\right] - \tau KL\left(\pi \| \pi_{ref}\right). \quad (5)$$

We provide a proof of this equivalence in Appendix D. We note that replacing $\pi_{\text{mix}}(\pi_{\text{sample}}, \pi_{\texttt{ref}})$ with $\pi_{\texttt{ref}}$ also gives an equivalent rewriting for the pessimistic Nash with no support restrictions (2).

3.2 P3O: An Efficient Implementation

The objectives for both the pessimistic game in Eq. (2) and the restricted version in (5) are challenging to implement for a number of reasons. In (2), the joint minimization over p and π' presents a challenging optimization landscape. Removing the explicit optimization over π' in Lemma 1 simplifies the inner minimization to only have one variable, but at the cost of changing the objective to have a more complicated log-partition function term. Consequently, we can no longer find stochastic gradient of the objective from a mini-batch of data, due to the non-linearity of the logarithm outside expectation.

To obtain a practical algorithm, we leverage ideas from variational inference [Jordan et al., 1999] and approximate the log-partition function with the expectation under a proposal distribution. Doing so, we obtain the following result, proved in Appendix E.

Lemma 2. *For any policies π and $\bar{\pi}$:*

$$\min_{p\in\mathcal{P}} -\log \mathbb{E}_{y\sim\pi_{mix}(\pi_{\texttt{ref}},\pi_{sample})}\left[\exp\left(\frac{-p(\pi,y)}{2\tau}\right)\right] \leq \min_{p\in\mathcal{P}} \mathbb{E}_{y\sim\pi_{mix}(\bar{\pi},\pi_{sample})}\left[\frac{p(\pi,y)}{2\tau}\right] + C,$$

where C is independent of the optimization variables π and p.

Given such an upper bound, we now design P3O to maximize the upper bound from Lemma 2 to find good policy π. We note that since the approximation of Lemma 2 is an upper bound on the true objective, maximizing the two is not equivalent, and the approximate objective simply takes the form of optimizing preference against the comparator $\bar{\pi}$ restricted to the sampling distribution. In the experiments, we choose $\bar{\pi}$ to be an exponentially moving average of past policy iterates. As a final step, we replace the constrained optimization over $\mathcal{P}$ to an unconstrained optimization over all preference functions in some parametric family by adding an additional loss term $-\mathcal{L}_{\text{pref}}(p, \mathcal{D})$, corresponding to the Lagrangian form of the constraint defining $\mathcal{P}$. The final objective function is

$$J(\pi, p) \doteq \mathbb{E}_{y\sim\pi_{\text{mix}}(\bar{\pi},\pi_{\text{sample}})}[p(\pi, y)] - \tau\text{KL}\left(\pi\|\pi_{\texttt{ref}}\right) - \lambda\mathcal{L}_{\text{pref}}(p, \mathcal{D}), \tag{6}$$

where we rescaled the objective to absorb the $1/2\tau$ on p into the corresponding hyper-parameters of KL and likelihood loss parameters (i.e., τ, λ). The resulting algorithm is shown in Algorithm 1.

4 Experimental results

To demonstrate the effectiveness of P3O in mitigating preference hacking, we compare it against existing preference optimization methods on the popular TL;DR summarization benchmark [Völske et al., 2017, Stiennon et al., 2020]. Following the setup in prior work on reward hacking [Eisenstein et al., 2024], we train the MLE preference model p_{MLE} by fine-tuning a T5 XL (3B) model [Raffel et al., 2020, Roberts et al., 2023]. The initial policy $\pi_{\texttt{ref}}$ is obtained by supervised fine-tuning of a T5 large model (770M) on the human summaries in the TL;DR dataset. Choosing a larger preference model than the polict is a commonly employed strategy for mitigating hacking [Eisenstein et al., 2024]. We initialize the training preference model $p_1 \doteq p_{\text{MLE}}$ as the MLE model.

For baselines, we consider existing preference-based RL methods, i.e., **NashEMA** [Munos et al., 2023] and **NashEMA restricted** as they correspond to P3O without and with support restrictions respectively, but without updating the preference function (i.e., $\eta_p = 0$). That is, they exactly match P3O without the pessimistic component. We refer to our methods as **P3O restricted** and **P3O**, corresponding to the versions with and without support restrictions. We train the policy for 20k steps, where every 2k steps we evaluate 100 samples generated by the policy against $\pi_{\texttt{ref}}$ using Gemini 1.0 Ultra [Team et al., 2023] as the judge. Details of the evaluation setup are provided in Appendix F.

Results: Figure 1 presents the results, where we notice that without any pessimism (i.e., no preference updates), we soon start to be dispreferred against $\pi_{\texttt{ref}}$, i.e., around 2k and 6k steps for NashEMA and NashEMA restricted, respectively. In contrast, for both variants of P3O, we observe that performance does not degrade significantly and, in fact, seems to continuously improve in the non-restricted case. We also notice, particularly in Figure 3 in Appendix F, that P3O restricted is quite effective at preventing length hacking [Eisenstein et al., 2024, Singhal et al., 2023], as well as maintaining minimal KL divergence from $\pi_{\texttt{ref}}$. However, the restricted version does not improve in terms preference over reference policy compared to the unrestricted P3O, presumably because the restricted policy has a small divergence and greater similarity in summaries with $\pi_{\texttt{ref}}$ as we also observe in Figure 3.[3] In summary, we find that pessimism keeps the the policies from drifting too far out of distribution of the preference data, and consequently successfully improves quality of its responses throughout the training period, in correspondence with our theory.

Future work Our results motivate some natural next steps. In this initial study, we only tried $\alpha = 0, 0.5$. Presumably a finer search over good values can uncover interesting trade-offs and enable even more robust learning. It would be also interesting to evaluate the case of $\bar{\pi}_t = \pi_t$ in the non-pessimistic case, which would correspond to SPO [Swamy et al., 2024], however we do not expect significant differences here. We would also like to extend our evaluation to another task.

[3] We also note that π_{sample} and $\pi_{\texttt{ref}}$ coincide in this case, so we do not expect the restricted version to offer significant gains even in theory.

References

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete Problems in AI Safety, 2016.

D. Calandriello, D. Guo, R. Munos, M. Rowland, Y. Tang, B. A. Pires, P. H. Richemond, C. L. Lan, M. Valko, T. Liu, R. Joshi, Z. Zheng, and B. Piot. Human Alignment of Large Language Models through Online Preference Optimisation, 2024.

S. Cen, J. Mei, K. Goshvadi, H. Dai, T. Yang, S. Yang, D. Schuurmans, Y. Chi, and B. Dai. Value-Incentivized Preference Optimization: A Unified Approach to Online and Offline RLHF, 2024.

J. Chen and N. Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

T. Coste, U. Anwar, R. Kirk, and D. Krueger. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023.

Q. Cui and S. S. Du. When is Offline Two-Player Zero-Sum Markov Game Solvable?, 2022.

J. Eisenstein, C. Nagpal, A. Agarwal, A. Beirami, A. D'Amour, D. J. Dvijotham, A. Fisch, K. Heller, S. Pfohl, D. Ramachandran, P. Shaw, and J. Berant. Helping or Herding? Reward Model Ensembles Mitigate but do not Eliminate Reward Hacking, 2024.

A. Fisch, J. Eisenstein, V. Zayats, A. Agarwal, A. Beirami, C. Nagpal, P. Shaw, and J. Berant. Robust preference optimization through reward model distillation. *arXiv preprint arXiv:2405.19316*, 2024.

L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

S. Guo, B. Zhang, T. Liu, T. Liu, M. Khalman, F. Llinares, A. Rame, T. Mesnard, Y. Zhao, B. Piot, J. Ferret, and M. Blondel. Direct Language Model Alignment from Online AI Feedback, 2024.

A. Huang, W. Zhan, T. Xie, J. D. Lee, W. Sun, A. Krishnamurthy, and D. J. Foster. Correcting the mythos of kl-regularization: Direct alignment without overparameterization via chi-squared preference optimization. *arXiv preprint arXiv:2407.13399*, 2024a.

A. Huang, W. Zhan, T. Xie, J. D. Lee, W. Sun, A. Krishnamurthy, and D. J. Foster. Correcting the Mythos of KL-Regularization: Direct Alignment without Overoptimization via Chi-Squared Preference Optimization. https://arxiv.org/abs/2407.13399v2, 2024b.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

Z. Liu, M. Lu, S. Zhang, B. Liu, H. Guo, Y. Yang, J. Blanchet, and Z. Wang. Provably Mitigating Overoptimization in RLHF: Your SFT Loss is Implicitly an Adversarial Regularizer, 2024.

R. Munos, M. Valko, D. Calandriello, M. G. Azar, M. Rowland, Z. D. Guo, Y. Tang, M. Geist, T. Mesnard, A. Michi, M. Selvi, S. Girgin, N. Momchev, O. Bachem, D. J. Mankowitz, D. Precup, and B. Piot. Nash Learning from Human Feedback, 2023.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

A. Roberts, H. W. Chung, G. Mishra, A. Levskaya, J. Bradbury, D. Andor, S. Narang, B. Lester, C. Gaffney, A. Mohiuddin, et al. Scaling up models and data with t5x and seqio. *Journal of Machine Learning Research*, 24(377):1–8, 2023.

P. Singhal, T. Goyal, J. Xu, and G. Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.

N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

G. Swamy, C. Dann, R. Kidambi, Z. S. Wu, and A. Agarwal. A Minimaximalist Approach to Reinforcement Learning from Human Feedback, 2024.

G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

M. Völske, M. Potthast, S. Syed, and B. Stein. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, 2017.

T. Xie, C.-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

C. Ye, W. Xiong, Y. Zhang, N. Jiang, and T. Zhang. Online Iterative Reinforcement Learning from Human Feedback with General Preference Model, 2024.

T. Zhang. From ε-entropy to kl-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, pages 2180–2210, 2006.

A Optimal Solution of the Example in Figure 2

Let $\pi_i = \pi(y_i)$ and $p_{23} = p(y_2, y_3)$ and $p_{13} = p(y_1, y_3)$. Then we can write the objective in this example as

$$\begin{aligned} V^\star &= \max_{\pi} \min_{\pi'} \min_{p \in \mathcal{P}} J(\pi, \pi', p) \\ \text{where } J(\pi, \pi', p) = p(\pi, \pi') = \big(& 0.5\pi_1\pi'_1 + 1 \cdot \pi_1\pi'_2 + p_{13}\pi_1\pi'_3 \\ & + 0 \cdot \pi_2\pi'_1 + 0.5\pi_2\pi'_2 + p_{23}\pi_2\pi'_3 \\ & + (1 - p_{13})\pi_3\pi'_1 + (1 - p_{23})\pi_3\pi'_2 + 0.5\pi_3\pi'_3\big). \end{aligned}$$

Upper-bound on $V^\star$: First note that for any π:

$$\min_{\pi'} \min_{p \in \mathcal{P}} J(\pi, \pi', p) \leq 0.5\pi_3 \qquad \text{and} \qquad \min_{\pi'} \min_{p \in \mathcal{P}} J(\pi, \pi', p) \leq 0.5\pi_1.$$

The first inequality follows by considering the choice $p_{13} = p_{23} = 1$, and the second inequality from considering the choice $\pi'_1 = 1$ and $p_{13} = 1$. Since both bounds hold simultaneously and $\pi_1 + \pi_3 \leq 1$, we can conclude that

$$V^\star \leq 0.25.$$

Lower-bound on $V^\star$: Choosing $\pi_1 = \pi_3 = 0.5$, we see that the objective value can be written as

$$\begin{aligned} J(\pi, \pi', p) = & \big(0.25\pi'_1 + 0.5\pi'_2 + 0.5p_{13}\pi'_3 \\ & + 0.5(1 - p_{13})\pi'_1 + 0.5(1 - p_{23})\pi'_2 + 0.25\pi'_3\big). \end{aligned}$$

First we observe that the minimum of this quantity is always attained at $p_{23} = 1$ and thus we can ignore the penultimate term. Consider now two cases:

- Case $\pi'_1 \leq \pi'_3$: Then the coefficient of p_{13} is non-negative and the minimum is attained at $p_{13} = 0$. This allows us to simplify the expression further as

  $$\begin{aligned} \min_{\pi'} \min_{p \in \mathcal{P}} J(\pi, \pi', p) &= \min_{\pi'} 0.25\pi'_1 + 0.5\pi'_2 + 0.5\pi'_1 + 0.5\pi'_2 + 0.25\pi'_3 \\ &= \min_{\pi'} 0.75\pi'_1 + \pi'_2 + 0.25\pi'_3 \\ &= 0.25 \end{aligned}$$

  where we choose $\pi'_3 = 1$ in the last step.

- Case $\pi'_1 \geq \pi'_3$: Then the coefficient of p_{13} is non-positive and the minimum is attained at $p_{13} = 1$. This gives

  $$\begin{aligned} \min_{\pi'} \min_{p \in \mathcal{P}} J(\pi, \pi', p) &= \min_{\pi'} 0.25\pi'_1 + 0.5\pi'_2 + 0.5\pi'_3 + 0.5\pi'_2 + 0.25\pi'_3 \\ &= \min_{\pi'} \min_{P \in \mathcal{P}} 0.25\pi'_1 + \pi'_2 + 0.75\pi'_3 \\ &= 0.25 \end{aligned}$$

  where the optimal solution is to choose $\pi'_1 = 1$.

Combining both cases, we can conclude that

$$V^\star \geq 0.25.$$

Optimal solution. Combining both upper- and lower-bounds, we can conclude that $V^\star = 0.25$ which is attained at $\pi_1 = \pi_3 = 0.5$.

B Definition and Analysis of Restricted Nash Policy

Recall that our preference dataset consists of tuples $(\boldsymbol{y}_{\mathrm{W}}, \boldsymbol{y}_{\mathrm{L}}) \overset{\text{i.i.d.}}{\sim} \pi_{\text{sample}}$. Based on existing statistical analysis of maximum likelihood estimation [Zhang, 2006], we expect that the maximum likelihood estimator $\hat{p}$ for the ground-truth preference model $p^\star$ satisfies:

$$\mathbb{E}_{\boldsymbol{y}_{\mathrm{W}}, \boldsymbol{y}_{\mathrm{L}} \overset{\text{i.i.d.}}{\sim} \pi_{\text{sample}}} |\hat{p}(\boldsymbol{y}_{\mathrm{W}}, \boldsymbol{y}_{\mathrm{L}}) - p^\star(\boldsymbol{y}_{\mathrm{W}}, \boldsymbol{y}_{\mathrm{L}})| \leq \varepsilon,$$

where we expect ε to scale as $O(\sqrt{\ln |\mathcal{P}|/|\mathcal{D}|})$, when learning from a finite preference model class $\mathcal{P}$. Given this, a natural definition for a policy π to be well-aligned with our sampling policy π_{sample} is that preference models which are close under π_{sample} should also be close under π. More formally, given a policy class Π and a sampling policy π_{sample}, and some constant $C \geq 1$, we define $\Pi(\pi_{\text{sample}}, C) \subseteq \Pi$ to be the set of policies such that for any $\pi \in \Pi(\pi_{\text{sample}}, C)$, we have:

$$\mathbb{E}_{y, y' \overset{\text{i.i.d.}}{\sim} \pi} |p_1(y, y') - p_2(y, y')| \leq C \cdot \mathbb{E}_{y, y' \overset{\text{i.i.d.}}{\sim} \pi_{\text{sample}}} |p_1(y, y') - p_2(y, y')|, \tag{7}$$

for any $p_1, p_2 \in \mathcal{P}$. We now demonstrate the effectiveness of this definition by showing that whenever we restrict the opponent policy π' to stay within $\Pi(\pi_{\text{sample}}, C)$, then the learned policy π enjoys strong learning guarantees under the (unknown) ground-truth preference model $p^\star$. For simplicity, we carry out the analysis without a KL term on the policy to $\pi_{\texttt{ref}}$, and our conclusions readily extend to the KL regularized objective.

Lemma 3 (Guarantee for restricted pessimistic Nash policy)**.** *We denote by $\hat{\pi}$ and $\pi^\star$ the restricted pessimistic Nash policy and restricted Nash policy respectively, that is*

$$\hat{\pi} = \arg\max_{\pi} \min_{p \in \mathcal{P}} \min_{\pi' \in \Pi(\pi_{\textit{sample}}, C)} p(\pi, \pi') \quad \textit{and} \quad \pi^\star = \arg\max_{\pi \in \Pi(\pi_{\textit{sample}}, C)} \min_{\pi' \in \Pi(\pi_{\textit{sample}}, C)} p^\star(\pi, \pi').$$

Then we have for any $\pi \in \Pi(\pi_{\textit{sample}}, C)$:

$$p^\star(\hat{\pi}, \pi) \geq \frac{1}{2} - C\varepsilon.$$

Proof. We start by noting that $\pi^\star$ is solving an anti-symmetric two player zero-sum game, and the constraint set $\Pi(\pi_{\text{sample}}, C)$ is an intersection of Π with constraints linear in π, so that it is convex set whenever π is convex. Hence we have that $\pi^\star \in \arg\min_{\pi \in \Pi(\pi_{\text{sample}}, C)} p^\star(\pi^\star, \pi)$ and $p^\star(\pi^\star, \pi^\star) = 0.5$. Let $\hat{\pi}' \in \arg\min_{\pi \in \Pi(\pi_{\text{sample}}, C)} \min_{p \in \mathcal{P}} p(\hat{\pi}, \pi)$. Then we have by definition:

$$\begin{aligned}
p^\star(\hat{\pi}, \pi) =& p^\star(\hat{\pi}, \pi) - p^\star(\pi^\star, \pi^\star) + 0.5 \\
\geq& \min_{p \in \mathcal{P}} p(\hat{\pi}, \hat{\pi}') - p^\star(\pi^\star, \pi^\star) + 0.5 \\
\geq& \min_{p \in \mathcal{P}} \min_{\pi' \in \Pi(\pi_{\text{sample}}, C)} p(\pi^\star, \pi') - p^\star(\pi^\star, \pi^\star) + 0.5,
\end{aligned}$$

where the first inequality is due to the definition of $\hat{\pi}'$, and the second follows from the definition of $\hat{\pi}$. Let $\widetilde{\pi} \in \arg\min_{\pi' \in \Pi(\pi_{\text{sample}}, C)} \min_{p \in \mathcal{P}} p(\pi^\star, \pi')$. Then we can further write

$$\begin{aligned}
p^\star(\hat{\pi}, \pi) \geq& \min_{p \in \mathcal{P}} p(\pi^\star, \widetilde{\pi}) - p^\star(\pi^\star, \widetilde{\pi}) + 0.5 \\
\geq& 0.5 - C\varepsilon,
\end{aligned}$$

where the first inequality is due to $\pi^\star \in \arg\min_{\pi' \in \Pi(\pi_{\text{sample}}, C)} p^\star(\pi^\star, \pi')$, and the second inequality follows from Equation 7. □

C Comparison with the unrestricted Nash

For ease of comparison with the analysis of Cui and Du [2022] for the unrestricted Nash case, let us consider a setting where $\mathcal{Y}$ is finite and there are no contexts. The set $\mathcal{P}$ consists of all possible

preference functions $p \in \mathcal{Y} \times \mathcal{Y} \to [0,1]$, such that $p(y,y') + p(y',y) = 1$. In this case, we see that the set $\Pi(\pi_{\text{sample}}, C)$ reduces to policies π such that for all y: $\pi(y)/\pi_{\text{sample}}(y) \leq C$. In particular, $\pi \in \Pi(\pi_{\text{sample}}, C)$ places no mass outside the support of π_{sample}. For any such policy, our returned policy $\hat{\pi}$ is preferred under $p^\star$ up to an error of $C\varepsilon$. As a particular corollary, since $\pi_{\text{sample}} \in \Pi(\pi_{\text{sample}}, C)$, we have that π is always preferred to the data collection policy.

In contrast, the situation is a bit different in the case of unrestricted Nash. Suppose there is a y_0 such that $\pi_{\text{sample}}(y_0) = 0$. Then even π_{sample} does not satisfy unilateral concentrability as defined in Cui and Du [2022]. In fact, no policy can satisfy unilateral concentrability in this case, and we get a vacuous guarantee out of their analysis.

In some sense, the contrast between our result from Lemma 3 and those of Cui and Du [2022] is analogous to the classical analysis of offline RL methods (see e.g. [Chen and Jiang, 2019]) and pessimistic offline RL techniques [Xie et al., 2021]. Without pessimism in offline RL, we end up with vacuous guarantees, while the pessimistic results allow a non-trivial sub-optimality bound against any policy well covered by the data collection policy. Similarly, the results of Cui and Du [2022] offer a strong guarantee when the data collection policy is sufficiently exploratory, but are rendered vacuous without this. In contrast, our analysis of the restricted Nash estimator offers an opportunistic guarantee, where we are able to adaptively compete with all policies which are well covered by the sampling policy.

D Proof of Lemma 1

Proof. We consider the following objective for $\alpha \in [0,1]$ and $\tau, \tau' \in \mathbb{R}^+$

$$\max_{\pi} \min_{p \in \mathcal{P}} \min_{\pi'} p(\pi, \pi') - \tau \mathrm{KL}\left(\pi \| \pi_{\texttt{ref}}\right) + \tau' \alpha \mathrm{KL}\left(\pi' \| \pi_{\texttt{ref}}\right) + \tau'(1-\alpha)\mathrm{KL}\left(\pi' \| \pi_{\text{sample}}\right) \tag{8}$$

Only looking at the inner minimization of π', we get

$$\min_{\pi'} p(\pi, \pi') + \tau' \alpha \sum_y \pi'(y) \log \frac{\pi'(y)}{\pi_{\texttt{ref}}(y)} + \tau'(1-\alpha) \sum_y \pi'(y) \log \frac{\pi'(y)}{\pi_{\text{sample}}(y)} \tag{9}$$

$$\min_{\pi'} p(\pi, \pi') + \tau \sum_y \pi'(y) \log \frac{\pi'(y)}{\pi_{\texttt{ref}}(y)^\alpha \pi_{\text{sample}}(y)^{1-\alpha}} \tag{10}$$

and thus, the optimal solution for π' can be written as

$$\pi'_\star(y) = \frac{1}{Z} \pi_{\texttt{ref}}(y)^\alpha \pi_{\text{sample}}(y)^{1-\alpha} \exp\left(-\frac{1}{\tau'} p(\pi, y)\right), \tag{11}$$

with partition function Z. Plugging this back in the objective above gives

$$\max_{\pi} \min_{p \in \mathcal{P}} p(\pi, \pi') - \tau \mathrm{KL}\left(\pi \| \pi_{\texttt{ref}}\right) + \tau' \sum_y \pi'_\star(y) \log \frac{\pi'_\star(y)}{\pi_{\texttt{ref}}(y)^\alpha \pi_{\text{sample}}(y)^{1-\alpha}} \tag{12}$$

$$= \max_{\pi} \min_{p \in \mathcal{P}} -\tau' \mathrm{KL}\left(\pi \| \pi_{\texttt{ref}}\right) - \tau' \log Z \tag{13}$$

$$= \max_{\pi} \min_{p \in \mathcal{P}} -\tau \mathrm{KL}\left(\pi \| \pi_{\texttt{ref}}\right) - \tau' \log \sum_y \pi_{\texttt{ref}}(y)^\alpha \pi_{\text{sample}}(y)^{1-\alpha} \exp\left(-\frac{1}{\tau'} p(\pi, y)\right) \tag{14}$$

$$= \max_{\pi} \min_{p \in \mathcal{P}} -\tau \mathrm{KL}\left(\pi \| \pi_{\texttt{ref}}\right) - \tau' \log \mathbb{E}_{y \sim \pi^\alpha_{\text{mix}}} \exp\left(-\frac{1}{\tau'} p(\pi, y)\right) + \tau' \log Z', \tag{15}$$

where $\pi^\alpha_{\text{mix}}(y) \propto \pi_{\texttt{ref}}(y)^\alpha \pi_{\text{sample}}(y)^{1-\alpha}$ and $Z' = \sum_y \pi_{\texttt{ref}}(y)^\alpha \pi_{\text{sample}}(y)^{1-\alpha}$ is a normalization constant, independent of optimization parameters. Dropping this term gives us an equivalent optimization objective in π. Setting $\alpha = 1/2$ and $\tau' = 2\tau$ completes the proof of the lemma. □

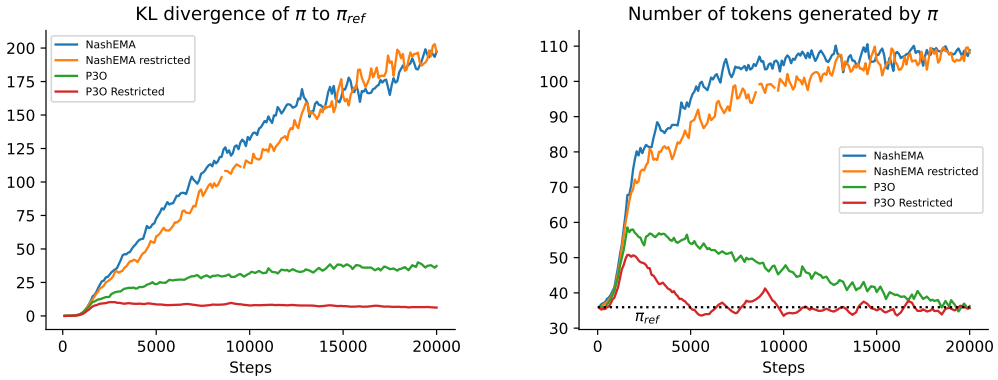

Figure 3: The horizontal axis in both plots represents the total number of learning steps. **(Left)** The figure plots the KL $(\pi_t \| \pi_{\texttt{ref}})$ on the vertical axis. Note that both P3O and P3O Restricted obtain much smaller divergence from $\pi_{\texttt{ref}}$, compared with the non-pessimistic variants. The divergence is particularly small for P3O Restricted, explaining its lower win-rate against $\pi_{\texttt{ref}}$ in Figure 1. **(Right)** The figure plots the length of summaries produced by π_t across different time steps. The dotted line represents the average length of summaries generated by $\pi_{\texttt{ref}}$. We note that while NashEMA and NashEMA Restricted exhibit length hacking, both variants of P3O depict minimal length hacking for longer periods of training. Prior works [Eisenstein et al., 2024] have observed long summaries to be a predominant mode of overoptimization on this dataset, and hence we can conclude that pessimism does effectively mitigate the issue.

E Proof of Lemma 2

Proof. Consider the log-sum-exp term with $\pi^{\alpha}_{\text{mix}}(y) = \frac{1}{Z'}\pi_{\texttt{ref}}(y)^{\alpha}\pi_{\text{sample}}(y)^{1-\alpha}$ and $Z' = \sum_y \pi_{\texttt{ref}}(y)^{\alpha}\pi_{\text{sample}}(y)^{1-\alpha}$ as

$$
\begin{aligned}
&= \log \mathbb{E}_{y\sim\pi^{\alpha}_{\text{mix}}} \exp\left(-\frac{1}{\tau'}p(\pi,y)\right) \\
&= \log \mathbb{E}_{y\sim\pi'}\left[\frac{\pi^{\alpha}_{\text{mix}}(y)}{\pi'(y)}\exp\left(-\frac{1}{\tau'}p(\pi,y)\right)\right] && (\pi' \text{ arbitrary}) \\
&\geq \mathbb{E}_{y\sim\pi'}\left[\log\left(\frac{\pi^{\alpha}_{\text{mix}}(y)}{\pi'(y)}\exp\left(-\frac{1}{\tau'}p(\pi,y)\right)\right)\right] && (\text{Jensen's inequality}) \\
&= -\frac{1}{\tau'}p(\pi,\pi') + \mathbb{E}_{y\sim\pi'}\log\left(\frac{\pi^{\alpha}_{\text{mix}}(y)}{\pi'(y)}\right) = -\frac{1}{\tau'}p(\pi,\pi') - \text{KL}\left(\pi' \| \pi^{\alpha}_{\text{mix}}\right).
\end{aligned}
$$

Setting $\alpha = 1/2$, $\tau' = 2\tau$ and taking the minimum over $p \in \mathcal{P}$ yields

$$
\min_{p\in\mathcal{P}} -\log \mathbb{E}_{y\sim\pi_{\text{mix}}(\pi_{\texttt{ref}},\pi_{\text{sample}})}\left[\exp\left(\frac{-p(\pi,y)}{2\tau}\right)\right] \leq \min_{p\in\mathcal{P}} \frac{p(\pi,\pi')}{2\tau} - \text{KL}\left(\pi' \| \pi_{\text{mix}}(\pi_{\texttt{ref}},\pi_{\text{sample}})\right). \tag{16}
$$

Choosing $\pi' = \pi_{\text{mix}}(\bar{\pi},\pi_{\text{sample}})$ gives the desired result with $C = -\text{KL}\left(\pi_{\text{mix}}(\bar{\pi},\pi_{\text{sample}}) \| \pi_{\text{mix}}(\pi_{\texttt{ref}},\pi_{\text{sample}})\right)$. □

F Experiments

F.1 Additional Results

Figure 3 plots the KL divergence and the length of summaries produced by the policy at different stages of learning. We note that for the same τ settings, NashEMA and NashEMA Restricted can move quite far away from $\pi_{\texttt{ref}}$ **(Left)** and also exhibit length hacking **(Right)**. In contrast, P3O is quite effective at staying close to $\pi_{\texttt{ref}}$ while also improving and avoiding length hacking, especially in the P3O Restricted cases, where policies produce summaries close to $\pi_{\texttt{ref}}$ while also improving in preference (see Figure 1).

F.2 Detailed setup of the empirical evaluation

Hyper-parameters: Policy is trained for $20{,}000$ steps. The learning rate for policy updates is fixed at $\eta_\pi = 10^{-5}$, and the preference learning rate is searched between $\eta_p \in \{2.5 \times 10^{-5}, 5 \times 10^{-5}\}$. We fix $\tau = 10^{-5}$ and sweep $\lambda \in \{1, 2, 4, 8, 16, 32, 64\}$ (6). We set the exponentially moving average (EMA) parameter to $\gamma = 0.0025$, and the set the context-length of input at 1024, whereas the generation length is set to 128.

Evaluation: We save a checkpoint for policies at every $2{,}000$ steps, and generate summaries from π_t for evaluation. To evaluate the learned model, we query Gemini 1.0 Ultra [Team et al., 2023] to judge which summary is better for the given input context. The prompt for evaluation is as follows:

> You are an expert summary rater who prefers very short and high quality summaries. Given a document and two candidate summaries, say 1 if SUMMARY1 is very short and high quality, and 2 if SUMMARY2 otherwise. Give a short reasoning for your answer.
> ARTICLE: <article-here>
> SUMMARY1: <summary-by-π_t>
> SUMMARY2: <summary-by-$\pi_{\texttt{ref}}$>.

To avoid any positional bias, we make two queries for each comparison, where we swap the order of SUMMARY1 and SUMMARY2, and average out the wins over 100 generations to get an estimate of the win-rate.