Reward Distillation Through Ratio Matching

Kenan Hasanaliyev [1 2] Schwinn Saereesitthipitak [1] Rohan Sanda [1]

Abstract

While Direct Preference Optimization (DPO) revolutionized language model alignment by eliminating the need for explicit reward models and reinforcement learning, scenarios with access to high-quality reward models (RMs) trained on extensive preference datasets still benefit from leveraging these resources. Reward model distillation techniques such as REBEL have emerged as part of a class of approaches that do not require the added complexity of reinforcement learning. In this paper, we show that REBEL can be derived as a ratio-matching objective with respect to DPO's optimal policy. In addition, we generalize ratio matching into distribution matching, formulating a new, principled alignment objective in the multi-completion setting where Group Relative Policy Optimization (GRPO) is commonly used.

1. Introduction

Direct Preference Optimization (DPO) (Rafailov et al., 2024) and its extensions have proven highly effective for aligning models using a preference dataset. In scenarios where there exists a pretrained policy π_θ and direct access to a reward model $r(x, y)$, Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022) remains the de facto standard. While DPO's key benefit lies in removing the need to train an explicit reward model, it can be useful to leverage reward models trained on many preference datasets to enhance alignment.

Reward model distillation techniques like the recently-proposed REBEL method (Gao et al., 2024) offer a middle ground by directly regressing policy likelihood ratios against reward differences, avoiding explicit reinforcement learning (RL) while still leveraging a learned reward signal. Despite their empirical promise, the connections among DPO, REBEL, and other distillation methods remain an underexplored area of research. Furthermore, in multi-completion settings where access to the reward scores of many completions is leveraged, the added complexity of RL is still required to an extent even in Group Relative Policy Optimization (GRPO) where the value function is bypassed (Zhihong Shao, 2024) but clipping and other stabilizations of Proximal Policy Optimization (Schulman et al., 2017) remain employed.

In this paper, we provide an alternative derivation of the REBEL objective, illustrating how it naturally arises from a ratio-matching formulation against DPO's optimal policy when the activation function is chosen to be logarithmic. In the multi-completion setting, where reward scores are computed for multiple model completions, we propose a principled extension that aligns the distribution of winning responses based on computed rewards with the distribution based on the model's implicit preferences, using KL divergence for distribution matching.

2. Related Work

The first dominant paradigm for aligning large language models (LLMs) with human preferences was Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022). The process consists of two stages: first, learning a reward model from human preference data, and second, fine-tuning the LLM policy using reinforcement learning algorithms, such as Proximal Policy Optimization (PPO) (Schulman et al., 2017), to maximize the learned reward. Although effective, RLHF can be unstable to train and computationally intensive because a separate reward model must be trained.

Reward-free Preference Optimization. DPO (Rafailov et al., 2024) is an offline supervised learning method that sidesteps the need for explicit reward model training by directly optimizing the policy using preference data. The key insight of DPO involves reparameterizing the Bradley-Terry preference model (Bradley & Terry, 1952) in terms of the policy being optimized and a reference policy, effectively showing that the language model itself implicitly defines a reward function. However, the Bradley-Terry assumption causes DPO to assign implicit rewards that tend towards infinite magnitude (Fisch et al., 2025) which can to lead to learning a degenerate policy. IPO (Azar et al., 2023)

[*]Equal contribution [1]Stanford University [2]Inception Labs. Correspondence to: Kenan Hasanaliyev <claserken@gmail.com>.

Proceedings of the ICML 2025 Workshop on Models of Human Feedback for AI Alignment, Vancouver, Canada. Copyright 2025 by the author(s).

was proposed as a more stable alternative to DPO when preference labels become (near-) deterministic. Rather than applying an unbounded log-odds transform to empirical preferences, IPO uses the raw preference scores directly and keeps them bounded. Consequently, IPO balances fitting the observed preferences and staying close the reference policy.

Using Explicit Reward Models in Offline Preference Optimization. REBEL (Reinforcement Learning via Regressing Relative Rewards) (Gao et al., 2024) was proposed as a simpler and more effective alternative to PPO, and can be viewed as a generalization of mirror descent and Natural Policy Gradient (Kakade, 2001). In REBEL, each policy update is simply a squared-error regression between the difference in log-likelihood-ratios of two candidate outputs to their reward difference as scored by an explicit, pretrained reward model. Although REBEL can utilize offline data, it can be run as an online RL algorithm as well. A similar method (Fisch et al., 2025) proposes an offline method that substitutes the DPO loss with a distillation loss that tries to match the reward differences between chosen and rejected outputs across the explicit, pretrained model and the implicit policy reward. In fact, REBEL (Gao et al., 2024) also arrives at the same loss function via a different perspective.

Ratio & Score Matching. Ratio matching (Hyvärinen, 2007; Sun et al., 2023) provides a method to learn unnormalized likelihood ratios directly from samples. Unlike standard maximum likelihood, ratio matching avoids directly modeling the full distribution and instead leverages a more flexible objective based on relative comparisons of unnormalized probabilities – making it well-suited to preference and distillation settings. Similarly, score matching methods (Meng et al., 2023) aim to estimate gradients of log-probabilities (scores) via methods like Fisher-divergence minimization without ever evaluating or normalizing probabilities.

In essence, this paper situates reward distillation techniques (like REBEL) within the landscape of preference optimization methods, highlighting its close relationship to DPO's theoretical underpinnings and IPO's objective structure. We unify the motivations behind REBEL (Gao et al., 2024) and reward model distillation (Fisch et al., 2025) by providing a novel ratio matching derivation for the REBEL loss, we offer an alternative theoretical justification for its effectiveness and clarify how it implicitly targets policy ratios consistent with an underlying reward model.

3. Preliminaries

3.1. DPO

Consider the original RLHF objective:

$$\max_{\pi} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi(\cdot|x)}[r(x,y) - \beta \mathrm{KL}(\pi(\cdot \mid x)||\pi_{\text{ref}}(x))] \quad (1)$$

The analytical solution for the optimal policy is the Boltzmann policy:

$$\pi^*(y \mid x) = \frac{1}{Z(x)}\,\pi_{\text{ref}}(y \mid x)\,\exp\left(\frac{r(x,y)}{\beta}\right). \quad (2)$$

where

$$Z(x) \;=\; \sum_{y} \pi_{\text{ref}}(y \mid x)\,\exp\Bigl(\frac{r(x,y)}{\beta}\Bigr)$$

is the (intractable) partition function and β is a regularization hyperparameter. Using this, we can write the reward function as:

$$r(x,y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} \;+\; \beta \log Z(x), \quad (3)$$

The main insight behind DPO is that Equation (2) and Equation (3) allow us to reparameterize the Bradley-Terry likelihood such that the intractable $Z(x)$ term is eliminated and the reward function does not need to be materialized explicitly:

$$p^*(y_1 \succ y_2 \mid x) = \Bigg[1 + \exp\Bigg(\beta \log \frac{\pi^*(y_2 \mid x)}{\pi_{\text{ref}}(y_2 \mid x)} - \beta \log \frac{\pi^*(y_1 \mid x)}{\pi_{\text{ref}}(y_1 \mid x)}\Bigg)\Bigg]^{-1} \quad (4)$$

This means we can optimize π_θ directly. Given a dataset of human preferences $\mathcal{D} = \{(x, y_w, y_\ell)\}$, DPO minimizes the loss

$$\mathcal{L}_{\text{DPO}}(\pi_\theta;\, \pi_{\text{ref}}) = -\,\mathbb{E}_{(x,y_w,y_\ell)\sim\mathcal{D}}\Big[\log\sigma\big(\beta \log \frac{\pi_\theta(y_w\,|\,x)}{\pi_{\text{ref}}(y_w\,|\,x)} - \beta \log \frac{\pi_\theta(y_\ell\,|\,x)}{\pi_{\text{ref}}(y_\ell\,|\,x)}\big)\Big], \quad (5)$$

where $\sigma(z) = 1/(1 + e^{-z})$.

3.2. IPO

While DPO uses a Bradley–Terry (log-odds) transform of empirical preferences, this transform becomes unbounded as preferences become (near-)determinisitc, causing DPO to collapse. To avoid this, IPO (Azar et al., 2023) simply regresses the policy's log-likelihood-ratio gap to a constant target.

Define the likelihood gap as

$$h_\pi(x, y_w, y_\ell) = \log \frac{\pi_\theta(y_w \mid x)\,\pi_{\text{ref}}(y_\ell \mid x)}{\pi_\theta(y_\ell \mid x)\,\pi_{\text{ref}}(y_w \mid x)}. \quad (6)$$

IPO minimizes

$$\mathcal{L}_{\mathrm{IPO}}(\theta) = \mathbb{E}_{(x,y_w,y_\ell)\sim\mathcal{D}}\Big[\big(h_\pi(x,y_w,y_\ell) \;-\; \tfrac{1}{2\beta}\big)^2\Big]. \quad (7)$$

By capping h_π in this way, IPO ensures that even when empirical preferences are extreme, the preference term remains bounded and the KL regularizer $\beta\,\mathrm{KL}(\pi_\theta \| \pi_{\mathrm{ref}})$ stays effective – preventing over-fitting and policy collapse. Note that the objective IPO optimizes for is entirely different from DPO.

3.3. REBEL

REBEL (Gao et al., 2024) starts from the mirror-descent update under a KL constraint:

$$\pi_{t+1}(y \mid x) \;\propto\; \pi_t(y \mid x)\,\exp\big(\eta\, r(x,y)\big), \qquad (8)$$

where π_t is the current policy at iteration t, $r(x,y)$ is a fixed, pretrained reward model, and $\eta > 0$ is a step-size. We can eliminate the partition function $Z_t(x)$ by comparing two samples (y, y') drawn from π_t. One can then fit a new parametric policy π_θ via a simple regression:

$$\mathcal{L}_{\mathrm{REBEL}}(\theta) = \mathbb{E}_{(x,y,y')\sim\mathcal{D}_t}\left[\left(\frac{1}{\eta}\big(\log\tfrac{\pi_\theta(y|x)}{\pi_{\theta_t}(y|x)} - \log\tfrac{\pi_\theta(y'|x)}{\pi_{\theta_t}(y'|x)}\big) - \big(r(x,y) - r(x,y')\big)\right)^2\right]. \qquad (9)$$

The REBEL loss equation shows that each REBEL update simply fits the change in log-likelihood ratios to the reward difference via squared-error regression. In other words, REBEL is regressing DPO's implicit reward gap against the true reward gap $r(x,y) - r(x,y')$.

3.4. GRPO

Group Relative Policy Optimization (GRPO) (Zhihong Shao, 2024) is an actor-only variant of PPO for multi-completion RLHF. For each prompt x, GRPO samples a group of G outputs from the current policy π_θ, scores each with a RM to obtain $\{r_i\}$, and computes the group-advantage:

$$A_i = \frac{r_i - \frac{1}{G}\sum_{j=1}^{G} r_j}{\mathrm{std}(\{r_j\})}\,.$$

It then maximizes the clipped objective averaged over the group:

$$J_{\mathrm{GRPO}}(\theta) = \mathbb{E}\Big[\tfrac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{|y_i|}\min\big(\rho_{i,t}\,\hat{A}_i,\; \mathrm{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)\,\hat{A}_i\big)\Big] - \beta\,\mathrm{KL}\big(\pi_\theta \| \pi_{\mathrm{ref}}\big). \qquad (10)$$

where $\rho_{i,t} = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\mathrm{old}}(y_{i,t}|x, y_{i,<t})}$

By using the group mean as a baseline, GRPO avoids the need for a separate value network, reducing memory and compute requirements while effectively aligning the policy across multiple candidate completions.

4. Ratio Matching Framework

We focus on the RLHF setting where we have full access to a reward model $r(x,y)$ as in REBEL. Our goal is to align our policy π_θ with the optimal policy π^* which can be written in terms of r in Equation (2). A direct KL-divergence minimization is intractable due to the partition function $Z(x)$. However, we can use a general ratio matching framework to bypass this issue.

4.1. Proposed Objective

We propose to match the ratios of policy probabilities under our model π_θ to the ratios under the optimal policy π^* as well as their reciprocals. To do this, we can define a loss $\mathcal{L}(\theta)$ that minimizes the difference between these ratios after applying a monotonic activation function $f(\cdot)$. A natural choice for this loss is the mean squared error over pairs of completions (y_1, y_2) for a given prompt x:

$$\mathbb{E}_{(x,y_1,y_2)\sim\mathcal{D}'}\left[\left(f\left(\frac{\pi^*(y_1|x)}{\pi^*(y_2|x)}\right) - f\left(\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)}\right)\right)^2 + \left(f\left(\frac{\pi^*(y_2|x)}{\pi^*(y_1|x)}\right) - f\left(\frac{\pi_\theta(y_2|x)}{\pi_\theta(y_1|x)}\right)\right)^2\right] \qquad (11)$$

where $\mathcal{D}'$ is a dataset of prompts and pairs of completions. The loss can be further simplified by substituting the definition of π^* from Equation (2) to get:

$$\frac{\pi^*(y_1|x)}{\pi^*(y_2|x)} = \frac{\pi_{\mathrm{ref}}(y_1|x)}{\pi_{\mathrm{ref}}(y_2|x)}\exp\left(\frac{r(x,y_1) - r(x,y_2)}{\beta}\right) \quad (12)$$

The function f can be any suitable monotone function to transform the ratios and determines the final form of the alignment objective.

4.2. Deriving REBEL

A particularly natural and convenient choice for the activation function is $f(z) = \log(z)$. By setting f to be the logarithm, our general loss in Equation 11 becomes:

$$\begin{aligned}
\mathcal{L}(\theta) &= \mathbb{E}\left[\left(\log\left(\frac{\pi^*(y_1|x)}{\pi^*(y_2|x)}\right) - \log\left(\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)}\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(\log\left(\frac{\pi_{\text{ref}}(y_1|x)}{\pi_{\text{ref}}(y_2|x)} e^{\frac{r(x,y_1)-r(x,y_2)}{\beta}}\right) - \log\left(\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)}\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(\log\frac{\pi_{\text{ref}}(y_1|x)}{\pi_{\text{ref}}(y_2|x)} + \frac{r(x,y_1)-r(x,y_2)}{\beta} - \log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)}\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{r(x,y_1)-r(x,y_2)}{\beta} - \left(\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)} - \log\frac{\pi_{\text{ref}}(y_1|x)}{\pi_{\text{ref}}(y_2|x)}\right)\right)^2\right]
\end{aligned}$$

where the reciprocals term in Equation (11) can be absorbed since $\log(x^{-1}) = -\log(x)$. The derived loss from our general ratio matching framework is equivalent to the REBEL loss objective in Equation (9) up to a constant factor.

4.3. Connection to IPO

Notice that the REBEL loss from Section 4.2 can be simplified as

$$\mathbb{E}\left[\left(\frac{r(x,y_1)-r(x,y_2)}{\beta} - h_\pi(y_1,y_2,x)\right)^2\right] \qquad (13)$$

where $h_\pi(y_1,y_2,x)$ is the log-likelihood ratio gap used in IPO:

$$h_\pi(y_1,y_2,x) = \log\left(\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)}\right) - \log\left(\frac{\pi_{\text{ref}}(y_1|x)}{\pi_{\text{ref}}(y_2|x)}\right). \qquad (14)$$

Recall that the IPO loss is given by $\mathbb{E}_{(y_w,y_l,x)\sim D}(h_\pi(y_w,y_l,x) - \frac{1}{2\beta})^2$, where the objective is to make the log-likelihood ratio gap h_π match a constant target. Hence, REBEL can be viewed as an instance of IPO, but instead of regressing h_π to a fixed target, the target is the regularized, reward-dependent difference $\frac{r(x,y_1)-r(x,y_2)}{\beta}$. This provides a stronger learning signal, as the model is encouraged to create a larger separation in policy probabilities for pairs with a large reward difference, and a smaller separation for pairs that are close in desirability. We note that Fisch et al. (2025) also highlight this connection between REBEL and IPO, but we reiterated it here for completeness.

5. Multi-Completion Extension

This work has primarily examined the application of a ratio-matching perspective given data with two completions y_1 and y_2 for prompt x. A natural generalization of this pairwise approach is to consider a setting with multiple completions for each prompt. That is, for each prompt x, G completions $y_1, y_2, \cdots, y_G$ along with their reward scores $r_1, r_2, \cdots, r_G$ will be available (which can be on-policy or off-policy). Here, r_i is shorthand for $r(x, y_i)$.

The multi-completion setting contains more reward scores from the reward model (RM), giving richer signal of the ideal response distribution to be targeted by the policy. The core idea is to align the distribution of the winning response over completions induced by the policy's implicit rewards with the distribution induced by the external reward model. Here, the implicit rewards are defined by $\hat{r_i} = \beta \log \frac{\pi_\theta(y_i|x)}{\pi_{\text{ref}}(y_i|x)}$ as stated in the DPO paper (Rafailov et al., 2024).

In Bradley-Terry reward modeling, the winning response is given by a softmax distribution over the rewards for responses of a fixed prompt. Hence, the ground-truth winning response distribution will be $\text{softmax}(r_1, \ldots, r_G)$, and the model's belief of the winning response distribution will be $\text{softmax}(\hat{r_1}, \ldots, \hat{r_G})$ where the regularization parameter β can be viewed as the temperature in the latter distribution. Hence, a natural choice for a multi-completion loss function to align the policy with the RM is to take the KL-divergence between the two discrete probability distributions:

$$\mathcal{L}_{\text{MC}} = \text{KL}\left(\text{softmax}(r_1, \ldots, r_G) \parallel \text{softmax}(\hat{r}_1, \ldots, \hat{r}_G)\right).$$

We summarize the multi-completion alignment procedure in Algorithm 1.

Algorithm 1 Multi-completion Alignment

```
For a given prompt $x$, sample $G$ completions: $\{y_1, y_2, \ldots, y_G\}$.
Obtain implicit rewards for each completion: $\{\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_G\}$ using
    $\hat{r_i} = \beta \log \frac{\pi_\theta(y_i|x)}{\pi_{\text{ref}}(y_i|x)}$
Obtain RM rewards for each completion: $\{r_1, r_2, \ldots, r_G\}$.
Calculate the loss $\mathcal{L}$ as the KL-divergence between the softmax of the RM rewards and the softmax of the implicit rewards:
    $\mathcal{L}_{\text{MC}} = \text{KL}\left(\text{softmax}(r_1, \ldots, r_G) \parallel \text{softmax}(\hat{r}_1, \ldots, \hat{r}_G)\right)$
```

We show that $\mathcal{L}_{\text{MC}}$ is both well-defined and minimized by the optimal policy π^* given in Equation (2).

Proposition 5.1. *$\mathcal{L}_{MC}$ is well-defined.*

Proof. The implicit rewards $\hat{r}_i$ are only defined up to a constant depending on x. However, the softmax distribution is shift-invariant, making $\mathcal{L}_{\text{MC}}$ well-defined. □

Proposition 5.2. *$\mathcal{L}_{MC}$ is minimized and equals 0 at the optimal policy π^*.*

Proof. KL-divergence is always non-negative, so it suffices to show that $\mathcal{L}_{\text{MC}} = 0$ when $\pi_\theta = \pi^*$. We have

$$\begin{aligned}\hat{r_i} &= \beta \log \frac{\pi^*(y_i \mid x)}{\pi_{\text{ref}}(y_i \mid x)} \\ &= \beta \log \frac{\frac{1}{Z(x)}\pi_{\text{ref}}(y_i \mid x) \exp\left(\frac{r(x,y_i)}{\beta}\right)}{\pi_{\text{ref}}(y_i \mid x)} \\ &= \beta \left(\frac{r(x,y_i)}{\beta} - \log Z(x)\right) \\ &= r(x,y_i) - \beta \log Z(x).\end{aligned}$$

The $\beta \log Z(x)$ term is a constant depending only on x. Hence, the softmax distribution of $\hat{r_i}$ and the softmax distribution of r_i are the same, making the KL divergence equal to 0. □

Proposition 5.1 and Proposition 5.2 together justify why $\mathcal{L}_{\text{MC}}$ is a theoretically principled choice of loss function. In our future work, we plan to include an experimental evaluation of Algorithm 1 on LLM alignment tasks for math and coding. Relevant baselines will include the currently popular GRPO method (Zhihong Shao, 2024). In contrast to Algorithm 1, GRPO requires additional hyperparameter tuning due to its PPO-style procedure and does not have a similar minimizer guarantee to Proposition 5.2.

6. Conclusion

Building on the foundations of DPO, we showed how REBEL and other offline reward model distillation techniques arise naturally from matching policy likelihood ratios to target ratios from either a reward model or constant preference gap. In addition, we have proposed an alignment objective serving as an extension to the ratio-matching framework that leverages multi-completion data, which has the potential to be more efficient, principled, and robust compared to existing approaches such as GRPO. Looking forward, we seek to explore alternative activation functions in ratio matching and perform empirical validation of our multi-completion extension.

References

Azar, M. G., Rowland, M., Piot, B., Guo, D., Calandriello, D., Valko, M., and Munos, R. A general theoretical paradigm to understand learning from human preferences, 2023. URL https://arxiv.org/abs/2310.12036.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2334029.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4302–4310, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Fisch, A., Eisenstein, J., Zayats, V., Agarwal, A., Beirami, A., Nagpal, C., Shaw, P., and Berant, J. Robust preference optimization through reward model distillation. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=E2zKNuwNDc.

Gao, Z., Chang, J. D., Zhan, W., Oertell, O., Swamy, G., Brantley, K., Joachims, T., Bagnell, J. A., Lee, J. D., and Sun, W. Rebel: Reinforcement learning via regressing relative rewards, 2024. URL https://arxiv.org/abs/2404.16767.

Hyvärinen, A. Some extensions of score matching. *Comput. Stat. Data Anal.*, 51:2499–2512, 2007. URL https://api.semanticscholar.org/CorpusID:2352990.

Kakade, S. M. A natural policy gradient. In Dietterich, T., Becker, S., and Ghahramani, Z. (eds.), *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/4b86abe48d358ecf194c56c69108433e-Paper.pdf.

Meng, C., Choi, K., Song, J., and Ermon, S. Concrete score matching: Generalized score matching for discrete data, 2023. URL https://arxiv.org/abs/2211.00802.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Sun, H., Yu, L., Dai, B., Schuurmans, D., and Dai, H. Score-based continuous-time discrete diffusion models, 2023. URL https://arxiv.org/abs/2211.16750.

Zhihong Shao, Peiyi Wang, Q. Z. R. X. J. S. M. Z. Y. L. Y. W. D. G. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.