Scaling Laws for Reward Model Overoptimization

Leo Gao
OpenAI

John Schulman
OpenAI

Jacob Hilton
OpenAI

Abstract

In reinforcement learning from human feedback, it is common to optimize against a reward model trained to predict human preferences. Because the reward model is an imperfect proxy, optimizing its value too much can hinder ground truth performance, in accordance with Goodhart's law. This effect has been frequently observed, but not carefully measured due to the expense of collecting human preference data. In this work, we use a synthetic setup in which a fixed "gold-standard" reward model plays the role of humans, providing labels used to train a proxy reward model. We study how the gold reward model score changes as we optimize against the proxy reward model using either reinforcement learning or best-of-n sampling. We find that this relationship follows a different functional form depending on the method of optimization, and that in both cases its coefficients scale smoothly with the number of reward model parameters. We also study the effect on this relationship of the size of the reward model dataset, the number of reward model and policy parameters, and the coefficient of the KL penalty added to the reward in the reinforcement learning setup. We explore the implications of these empirical results for theoretical considerations in AI alignment.

1 Introduction

Goodhart's law is an adage that states, "When a measure becomes a target, it ceases to be a good measure." In machine learning, this effect arises with proxy objectives provided by static learned models, such as discriminators and reward models. Optimizing too much against such a model eventually hinders the true objective, a phenomenon we refer to as *overoptimization*. It is important to understand the size of this effect and how it scales, in order to predict how much a learned model can be safely optimized against. Moreover, studying this effect empirically could aid in the development of theoretical models of Goodhart's law for neural networks, which could be critical for avoiding dangerous misalignment of future AI systems.

In this work, we study overoptimization in the context of large language models fine-tuned as reward models trained to predict which of two options a human will prefer. Such reward models have been used to train language models to perform a variety of complex tasks that are hard to judge automatically, including summarization [Stiennon et al., 2020], question-answering [Nakano et al., 2021, Menick et al., 2022], and general assistance [Ouyang et al., 2022, Bai et al., 2022, Glaese et al., 2022]. Typically, the reward model score is optimized using either policy gradient-based reinforcement learning or best-of-n sampling, also known as rejection sampling or reranking. Overoptimization can occur with both methods, and we study both to better understand whether and how overoptimization behaves differently across both methods.

A major challenge in studying overoptimization in this context is the expense of collecting human preference labels. A large number of labels are required to accurately estimate overall preference probabilities, and this is exacerbated by small effect sizes and the need to take many measurements in order to fit scaling laws. To overcome this, we use a synthetic setup that is described in Section 2, in which labels are supplied by a "gold-standard" reward model (RM) instead of humans.

Preprint. Under review.

Our main results are empirically validated functional forms for the gold reward model scores R as a function of the Kullback–Leibler divergence from the initial policy to the optimized policy $\mathrm{KL} := D_{\mathrm{KL}}\left(\pi \parallel \pi_{\mathrm{init}}\right)$, which depends on the method of optimization used. This KL distance between the initial and optimized policies increases monotonically during during RL training (fig. 14), and can be computed analytically as a function of n for BoN. Further, because it is a quadratic metric of distance [Bai et al., 2022, Section 4.3], we will define $d := \sqrt{D_{\mathrm{KL}}\left(\pi \parallel \pi_{\mathrm{init}}\right)}$, and write our functional forms in terms of d.

We find empirically that for best-of-n (BoN) sampling,

$$\boxed{R_{\mathrm{bo}n}\left(d\right) = d\left(\alpha_{\mathrm{bo}n} - \beta_{\mathrm{bo}n} d\right)},$$

and for reinforcement learning,[1]

$$\boxed{R_{\mathrm{RL}}\left(d\right) = d\left(\alpha_{\mathrm{RL}} - \beta_{\mathrm{RL}} \log d\right)},$$

Here, $R(0) := 0$ by definition and α_{RL}, β_{RL}, $\alpha_{\mathrm{bo}n}$ and $\beta_{\mathrm{bo}n}$ are parameters that may depend on the number of proxy reward model parameters, the size of the proxy reward model dataset, and so on. We see that these scaling laws make accurate predictions.

We also find the following.

- **RL versus best-of-n.** As a function of the KL divergence, reinforcement learning tends to be slower than best-of-n sampling at both optimization and overoptimization. This suggests inadequacies with using KL to compare amount of (over)optimization across methods. However, the relationship between the proxy reward model score and the gold reward model score is similar for both methods.
- **Smooth coefficient scaling.** The α and β coefficients in the BoN and RL functional forms vary smoothly with the number of proxy reward model parameters, following approximate logarithmic trends.[2] This allows prediction of attained gold RM score.
- **Weak dependence on policy size.** While larger policies perform better overall and benefit less from optimization against an RM as measured by increase in gold reward, they lead to very similar amounts of overoptimization, as measured through the gap between the proxy and gold scores (which indicates the shortfall between predicted and actual reward), and KL distance at which the maximum gold RM score is attained.
- **KL penalty ineffectiveness.** In our reinforcement learning setup, using a KL penalty increases the proxy reward model score that can be achieved for a given KL divergence, but this does not correspond to a measurable improvement in the gold RM score–$\mathrm{KL}_{\mathrm{RL}}$ frontier. However, we note this result could be particularly sensitive to hyperparameters.

Finally, we discuss the implications of these findings for Reinforcement Learning From Human Feedback (RLHF), existing models of Goodhart's law, and AI Alignment more broadly.

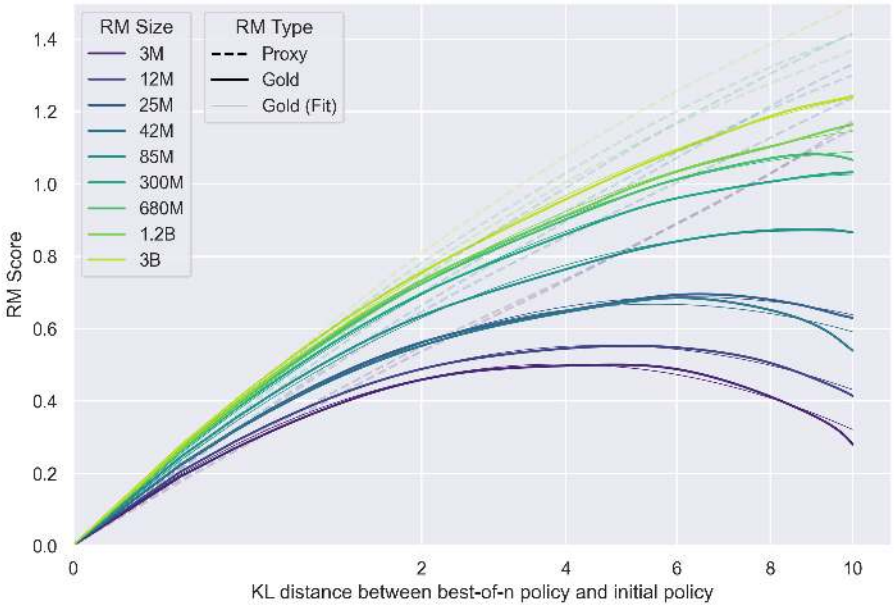
2 Methodology

The setting used throughout this paper is the same as for InstructGPT [Ouyang et al., 2022]. In our environment, the observations are text prompts and the policy is used to generate a response to the prompt. The prompts are drawn from a broad range of natural language instructions describing different language model tasks. Then, a learned RM is used to provide the reward signal for the response, which is used by either RL or BoN for optimization.

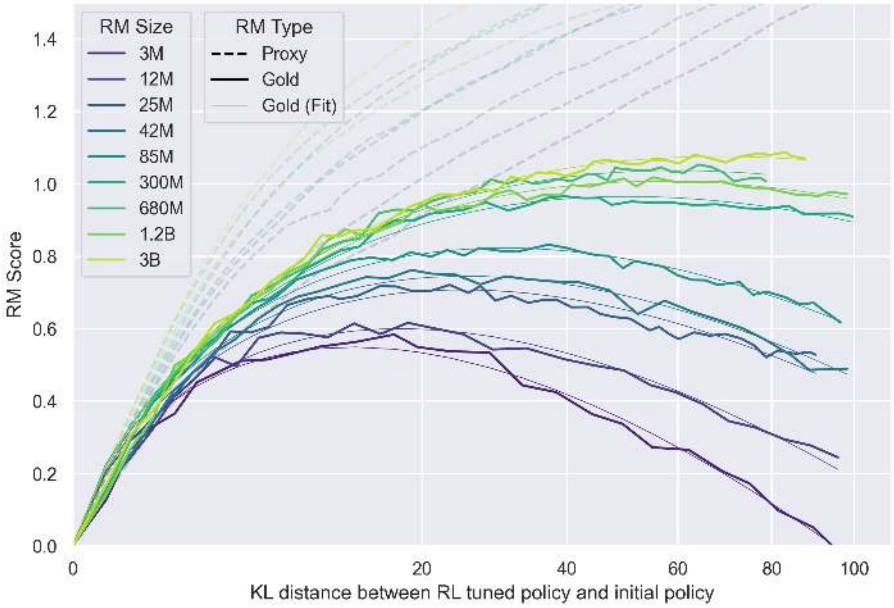
For all experiments, we use pretrained GPT-3 series language models as the initial checkpoint [Brown et al., 2020]. All initial policies are trained with supervised fine-tuning (SFT) on human-generated InstructGPT demonstrations [Ouyang et al., 2022] for 2 epochs. All RMs also use the GPT-3 architecture but have an added scalar head to output the reward.

[1] We note that this form likely does not hold near the origin, as it has infinite slope there. We experimented with a number of different forms, but found worse fits and extrapolation. See appendix B for more details.

[2] The coefficient α_{RL} in particular being nearly independent of RM parameter count.



(a) BoN



(b) RL

Figure 1: Reward model (RM) parameter size scaling experiments using the InstructGPT environment. Policy size is held constant (`1.2B`), while reward model size is varied. The x-axes have a square-root scale. Note that the plots have different x-axes. The gold reward represents the ground truth reward; we observe that when we optimize for a learned proxy of the gold reward, the gold reward initially increases and later decreases. We show that our functional forms fit this effect well.

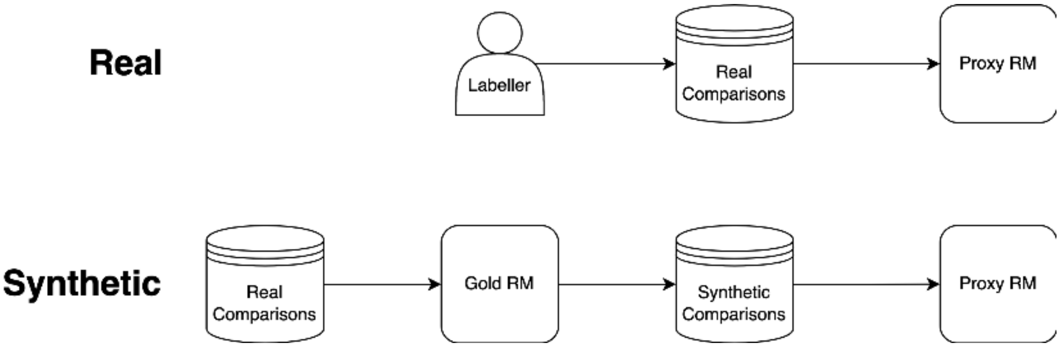


Figure 2: Diagram of the real and synthetic RM training setups. Human labellers generate comparison data. In the real RLHF setting, this data is used to train a proxy RM that is optimized by RL/BoN. In our synthetic setting, we instead use a "Gold RM" as our ground truth. In both settings, the proxy RM is a proxy for the ground truth process generating the labels (either the human or gold RM).

The RL experiments use Proximal Policy Optimization (PPO) [Schulman et al., 2017]. KL penalty for all RL experiments is set to 0 except for in section 3.6. See appendix C for all other hyperparameters. We mostly use defaults for the PPO hyperparameters; thus, it is possible that there exist different trends for other hyperparameter configurations.

In BoN, we generate n trajectories for the policy and use the reward model to pick the one with the highest proxy RM score. We use the unbiased estimator from Nakano et al. [2021, Appendix I] to compute all of the gold and proxy scores for intermediate n between 1 and the maximum n with lower variance and more efficiently than the naive estimator of randomly sampling n samples with replacement repeatedly and taking the mean of the maximum gold and proxy RM scores. The KL distances for BoN are computed analytically: $\mathrm{KL}_{\mathrm{bo}n} = \log n - \frac{n-1}{n}$ [Stiennon et al., 2020, Appendix G.3].

2.1 Synthetic Data Setup

Because getting a ground truth gold reward signal from human labellers is expensive, we instead use a synthetic task where the ground truth is defined to be the output of a particular large "gold" RM. The 6B reward model from Ouyang et al. [2022] is used as the gold RM, and our proxy RMs vary from 3M to 3B parameters[3]. This synthetic gold reward is used to label pairs of rollouts from the policy given the same prompt to create synthetic RM training data. The synthetic comparisons are created deterministically by always marking the trajectory with the higher gold RM score as preferred.[4] We generate 100,000 synthetic comparisons and reserve 10% of these as a held out test set for computing the validation loss of RMs.

See fig. 2 for a diagram of the synthetic setup.

2.2 Recalibration

The RM scores are translation-invariant, so to ensure comparability across different reward models, we recenter each RM such that the average reward of the initial policy is 0. We also unit normalize the variance of the gold RM scores.[5] Because our hard thresholding synthetic data setup produces labels that are miscalibrated (since they do not incorporate the gold RM's confidence), we recalibrate the proxy RMs by rescaling the logits to minimize cross-entropy loss using a validation set of soft labels. All renormalization and recalibration is applied after the experiments; this does not affect BoN at all, and likely has no impact on RL because Adam is loss scale invariant, though it is possible that there are slight differences due to algorithmic details.

[3] We originally trained two additional RMs smaller than 3M parameters, which achieved near-chance accuracy and were off-trend, and so were excluded.

[4] We had experimented with sampling for creating labels, but observed noisier results.

[5] We later decided this was unnecessary but decided not to change it.

3 Results

3.1 Fitting and validating functional forms

We chose our functional forms through experimentation with all RM data and parameter scaling curves in the remainder of this paper.

The BoN functional form was hypothesized using data up to $n = 1000$. In order to validate the functional forms, we performed a BoN experiment with up to $n = 60,000$ (KL ≈ 10 nats), after only having seen data up to $n = 1,000$ (KL ≈ 6 nats). As this experiment was conducted after the functional form was hypothesized based on data up to 6 nats, this was a true advance prediction.

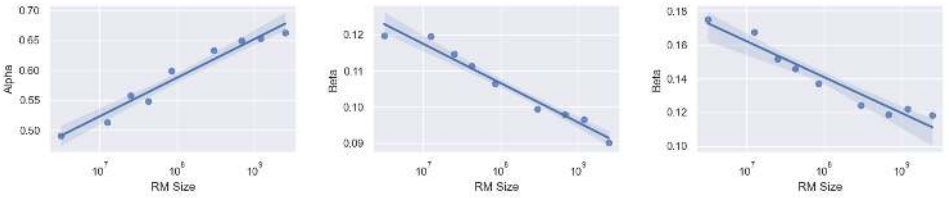
We also test extrapolation of the BoN and RL functional forms from low KLs to to unseen larger KLs; see fig. 26 for details.

We also attempted to model the proxy scores but were unable to obtain a satisfactory fit. For BoN, despite visual similarity, a linear fit ($d\alpha_{\text{bo}n}$) did not work well (fig. 20). The predictions for RL and BoN are not as easily modelled as the gold score predictions. We leave a better understanding of the proxy RM score behavior to future work.

3.2 Scaling with RM Parameter Count

We hold policy size (1.2B) and data size (90,000) constant (fig. 1). We observe that for the gold RM scores, $\alpha_{\text{bo}n}$ and $\beta_{\text{bo}n}$ change smoothly with RM size (figs. 3a and 3b). For RL, we find that we can hold α_{RL} constant across all RM sizes, resulting in a clean scaling curve for β_{RL} (fig. 3c). These scaling laws allow us to predict properties of training runs; for instance, we can also predict the peak gold RM scores for different RM sizes (fig. 12).

When modelled using the same functional forms as the respective gold scores, the proxy score fits have much lower values of $\beta_{\text{bo}n}$. We also see smooth scaling in the proxy score's $\alpha_{\text{bo}n}$ and $\beta_{\text{bo}n}$. However, for the reasons in section 3.1, we are less confident about these fits. For both BoN and RL, we observe systematic underestimates of the proxy reward model when extrapolated to higher KLs. Both appear to eventually grow roughly linearly in $\sqrt{\text{KL}}$, as in Bai et al. [2022].



(a) $\alpha_{\text{bo}n}$ (b) $\beta_{\text{bo}n}$ (c) β_{RL}

Figure 3: The values of $\alpha_{\text{bo}n}$, $\beta_{\text{bo}n}$ and β_{RL} in the BoN and RL overoptimization scaling laws for both proxy (dashed line) and gold (solid line) rewards as they scale with parameter count.

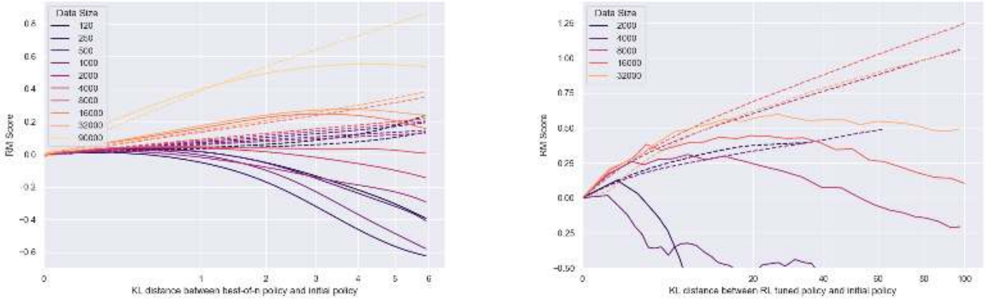
3.3 Scaling with RM Data Size

We hold RM size constant (12M) and sweep RM data size for both RL and BoN.[6]. Overall, the results are consistent with intuition: more data leads to better gold scores and less goodharting. The scaling of α and β with data size are not as cleanly described as for RM size scaling (fig. 17, fig. 18).

For all RM sizes, we observe that for amounts of data less than around 2,000 comparisons[7], there is very little improvement over near-chance loss (Figure 6). This is also reflected in gold scores after optimization (fig. 21). After this threshold, all models improve with more data, though larger RMs

[6] For BoN, we actually sweep all combinations of RM size and data size; see fig. 10. For a version of fig. 4a against a 3B RM, see fig. 19.

[7] To test the hypothesis that some minimum number of RM finetuning steps is needed, we control for the number of SGD steps by running multiple epochs and observe that running 4 epochs instead of 1 yields no change in gold score whatsoever, whereas 1 epoch of 4 times as much data performs substantially better (fig. 13).



(a) BoN

(b) RL

Figure 4: RM data scaling experiments. RM size is held constant (12M), while RM data is varied. The x-axis has a square root scale. Note that the plots have different axes. Dotted lines indicate proxy rewards, solid lines indicate gold rewards.

generally improve faster. Interestingly, although larger RMs result in better gold scores overall, they do not appear to have this critical threshold substantially earlier than smaller models.[8]

We hypothesized that two RMs of equal validation loss would achieve the same robustness against optimization, regardless of the combination of RM size and RM data size. Our results provide some weak evidence for this hypothesis (fig. 5).

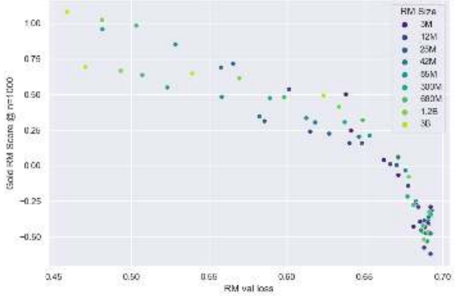


Figure 5: RM validation loss vs BoN RM score @ n=1000. Most points in this figure are already averaged over multiple seeds.

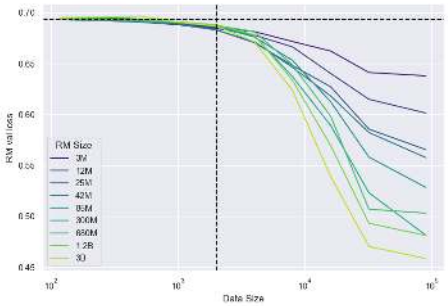


Figure 6: RM losses, broken down by data size and RM size

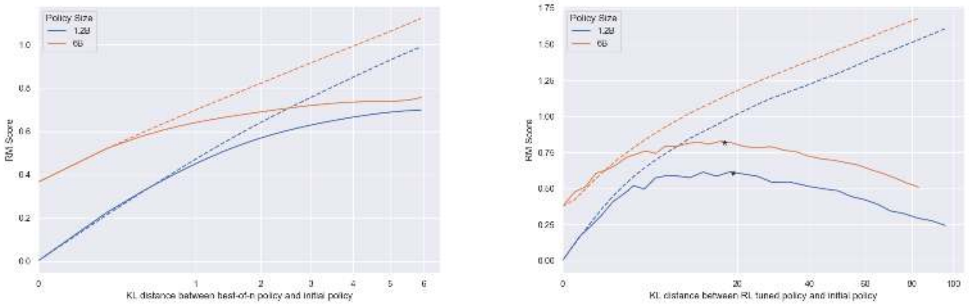
3.4 Scaling with Policy Size

We briefly explore the impact of policy size by holding the RM size constant (12M) and evaluating two different policy sizes. We also perform the same experiment with a different RM size (3B), observing similar results (fig. 22).

Larger policies see less benefit from optimization against an RM, but don't overoptimize more. We observe that the 6B policy run has a smaller difference between its initial and peak gold reward model scores than the 1.2B policy run. This is most visible in the BoN plot (fig. 7a).[9] However, while we might expect that a larger policy overoptimizes substantially faster, contrary to intuition, we find that both gold scores peak at almost the same KL. In fact, the gap between the proxy and gold scores is almost the same between the two policy sizes (fig. 24). We can interpret this gap, the shortfall

[8] This result contradicts some other internal findings; thus, it is possible that this is an artifact of this particular setup.

[9] For a version of the RL plot (fig. 7b) with all runs starting at 0, see fig. 23.



(a) BoN

(b) RL

Figure 7: Policy scaling experiments. RM size is held constant (12M), while policy size is varied. The x-axis has a square root scale. Note that the plots have different axes. Dotted lines indicate proxy rewards, solid lines indicate gold rewards. The asterisks in the RL plot indicate the max gold score for each policy size.

between the predicted and actual rewards, as being indicative of the extent to which the proxy RM is exploited. We discuss this result further in section 4.4.

3.5 RL vs BoN

A priori, we might expect reinforcement learning via PPO [Schulman et al., 2017] and best-of-n to apply optimization in very different ways. As such, we ask whether this difference in optimization results in different overoptimization characteristics. Similarities would potentially indicate candidates for further study in gaining a more fundamental understanding of overoptimization in general, and differences opportunities for better optimization algorithms. We note the following:

RL is far less KL-efficient than BoN. Viewing KL distance as a resource to be spent, we observe that RL "consumes" far more KL than BoN. This means that both optimization and overoptimization require more KL to occur with RL. Intuitively, BoN searches very locally around the initial policy, and thus $\mathrm{KL}_{\mathrm{bo}n}$ increases with roughly $\log(n)$. For RL on the other hand, each step modifies the policy from the policy of the previous step—KL increases approximately quadratically with step in the absence of KL penalty (Figure 16, Figure 14). An implication of this result is that KL distance is an inadequate metric for quantity of (over)optimization; we discuss this further in section 4.1.

When looking at proxy vs gold RM scores, BoN and RL look more similar. The proxy RM score is another possible metric for quantity of optimization, because it is the value that is being directly optimized for. Using it as the metric of optimization leads to significantly more analogy between RL and BoN than KL distance does. However, we do observe that RL initially has a larger proxy-gold gap (i.e requires more proxy RM increase to match BoN), but then peaks at a higher gold RM score than BoN (fig. 8).

3.6 Effect of KL Penalty

We observe in our setting that when varying the KL penalty for RL, the gold RM scores depend only on the KL distance of the policy $\mathrm{KL}_{\mathrm{RL}}$ (Figure 9). The KL penalty only causes the gold RM score to converge earlier, but does not affect the $\mathrm{KL}_{\mathrm{RL}}$-gold reward frontier, and so the effect of the penalty on the gold score is akin to early stopping (Figure 14). However, we have seen some evidence that this result could be particularly sensitive to hyperparameters.

Because we observe that using KL penalty has a strictly larger proxy-gold gap, we set KL penalty to 0 for all other RL experiments in this paper.

It is important to note that PPO's surrogate objective incorporates an implicit penalty on $D_{\mathrm{KL}}\left(\pi_{\mathrm{old}} \parallel \pi\right)$, where π_{old} is a recent policy (not the initial policy) [Schulman et al., 2017]. This penalty is used to control how fast the policy changes, but also has an indirect effect on the KL we

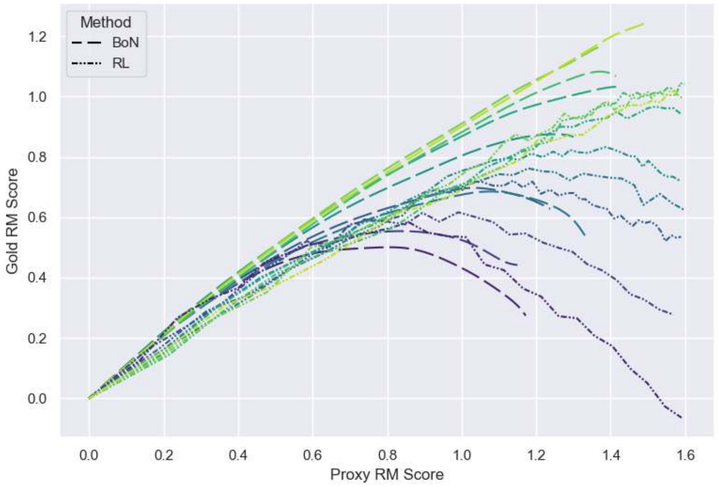


Figure 8: Proxy vs gold RM score for both BoN and RL. RL curves are truncated to a proxy RM score of 1.6 for readability.

study here, $D_{\mathrm{KL}}\left(\pi \parallel \pi_{\mathrm{init}}\right)$, causing it to grow much more slowly (providing the implementation is well-tuned). We do not know why this indirect effect appears to lead to less overoptimization than an explicit KL penalty.

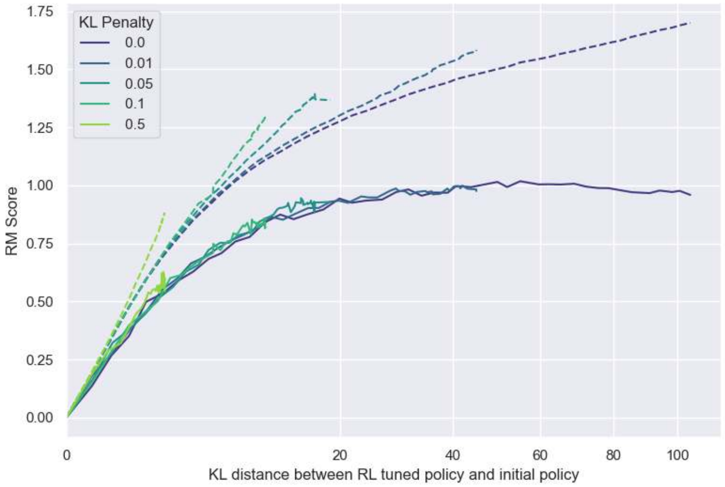


Figure 9: RL experiments with various KL penalties. Policy size (1.2B) and RM size (1.2B) are held constant. Dotted lines indicate proxy rewards, solid lines indicate gold rewards. We observe the effect of the KL penalty on the gold score as being equivalent to early stopping.

4 Discussion

4.1 KL as a measure of amount of optimization

For any given fixed optimization method, KL yields clean scaling trends, such as the ones observed in section 3.2, and consistent peak gold RM score KLs as in section 3.4. However, because it's clear that different methods of optimization spend KL very differently (section 3.5), it should not be used to compare the amount of optimization between different optimization algorithms. There exist pertubations to a policy that are orthogonal to the reward signal that would result in increases in KL that do not increase either gold or proxy reward; conversely, extremely small but well targeted perturbations could substantially change the behavior of the policy within a small KL budget.

4.2 Relation to Goodhart Taxonomy

One useful taxonomy for various Goodhart effects is presented in Manheim and Garrabrant [2018], categorizing Goodhart's Law into 4 categories: Regressional, Extremal, Causal, and Adversarial. In this section, we discuss our results in the framework of this taxonomy.

4.2.1 Regressional Goodhart

Regressional Goodhart occurs when our proxy RMs depend on features with noise. The simplest toy example of this is a proxy reward $\hat{X}$ which is exactly equal to the gold reward X plus some independent noise Z. When optimizing against this proxy, some amount of optimization power will go to selecting for noise, leading to a gold reward less than predicted by the proxy.

More formally, for independent absolutely continuous random variables X and Z with X normally distributed and either (a) Z normally distributed or (b) $|Z - \mathbb{E}[Z]| < \delta$ for some $\delta > 0$, this model predicts a gold reward that is:

$$\mathbb{E}[X \mid \hat{X} = \hat{x}] = \mathbb{E}[X] + (\hat{x} - \mathbb{E}[X] - \mathbb{E}[Z]) \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + \mathrm{Var}(Z)} + \varepsilon \tag{1}$$

where $\varepsilon = 0$ in case (a) and $\varepsilon = o(\mathrm{Var}(Z))$ as $\delta \to 0$ in case (b). See appendix A for the proof.

Intuitively, we can interpret eq. (1) as stating that the optimization power expended is divided between optimizing the gold reward and selecting on the noise proportional to their variances. This also implies that if this is the only kind of Goodhart present, the gold reward must always increase monotonically with the proxy reward; as we observe nonmonotonic behavior (fig. 8), there must be either noise distributions violating these assumptions or other kinds of Goodhart at play.

This result lends itself to an interpretation of the α term in the RL and BoN gold score scaling laws: since for both RL and BoN the proxy scores are roughly linear in $\sqrt{\mathrm{KL}}$, the difference in the slope of the proxy score and the linear component of the gold score (i.e the α term) can be interpreted as the amount of regressional Goodhart occurring.

4.2.2 Extremal Goodhart

We can think of out of distribution failures of the RM as an instance of extremal Goodhart. As we optimize against the proxy RM, the distribution of our samples shifts out of the training distribution of the RM, and thus the relation between the proxy and gold scores weakens. For instance, suppose in the training distribution a feature like answer length always indicates a higher quality answer, and thus the proxy RM infers that longer answers are always better, even though at some point outside the training distribution, selecting on longer answers no longer improves quality.[10]

We can also think of this as the proxy failing to depend on relevant features; this failure bears resemblance to the setting considered in Zhuang and Hadfield-Menell [2020], where a failure of the proxy to consider all features, under certain conditions, leads to overoptimization with unbounded loss of utility regardless of optimization method.

We expect extremal Goodharting to be primarily responsible for the nonmonotonicity of the gold RM scores in this paper, and is mostly responsible for the β term, which in the limit of optimization, results in an unbounded loss of utility. This lends a natural interpretation to the smooth decrease in β for both BoN and RL with increased RM size as smooth improvements in model robustness (fig. 3).

[10] Optimized policies producing very long answers even when a short answer would be preferred is a real issue that we have observed in other experiments in the InstructGPT setting.

4.2.3 Causal Goodhart

We can think of causal Goodhart as being a generalization of regressional Goodhart: there may exist correlations between features and gold score where the causal structure of the problem is such that selecting on the feature does not increase the gold score. For instance, suppose answer length is correlated with quality due to some other common cause (say, informativeness); then, the proxy RM may learn to use answer length as a feature, and when we select against the proxy we get longer answers that do not increase on actual quality.[11] In our experiments, we would observe causal Goodhart as behaving similarly to regressional Goodhart.

4.2.4 Adversarial Goodhart

Adversarial Goodhart occurs when the policy actively manipulates the proxy. We do not expect the effects of adversarial Goodhart to be captured in this work, as the models involved are not powerful enough to implement adversarial strategies. However, given the constant improvement of ML capabilities, it is entirely plausible that ML systems will one day become capable enough to do so [Hubinger et al., 2019]. When this occurs, the scaling laws observed in this paper may break down. Thus, we advise caution when using these results for extrapolation.

4.3 Implications for iterated RLHF

When conducting reinforcement learning from human feedback, it is preferable to use an online setup, in which fresh human feedback data is periodically used to train a new reward model, to mitigate overoptimization [Bai et al., 2022]. Our scaling law allows us to analyze the effect of this iterative approach under some simplifying assumptions. We assume firstly that the scaling coefficients α_{RL} and β_{RL} remain constant across iterations, and secondly that the distance $d = \sqrt{\text{KL}}$ is additive across iterations (because of how KL appears to grow empirically as in Figure 14). Under these assumptions, the final gold reward model score after k iterations each covering a distance d/k is given by

$$R_{\text{RL}}(d) = d\left(\alpha_{\text{RL}} - \beta_{\text{RL}} \log(d) + \beta_{\text{RL}} \log(k)\right).$$

Two interesting observations follow from this. Firstly, the iterative approach does not affect any Goodharting captured by the α_{RL} term (such as regressional Goodharting, as discussed in Section 4.2.1). Secondly, the effect of the iterative approach is to increase the final gold RM score by an amount proportional to both d and $\log(k)$, namely

$$\beta_{\text{RL}} d \log(k).$$

Note that this result can only hold up to some maximum value of k, and we expect our scaling law to break down below some minimum distance. Further research is required to determine what this minimum is, as well as to what extent our simplifying assumptions hold in practice.

4.4 Policy size independence

Our observation that larger SFT policies seem to exhibit the same amount of overoptimization during RL implies that larger policies do not increase the amount of optimization power applied to the RM or learn faster, even though they start out with higher performance on the gold score. While it is expected that larger policies have less to gain from optimizing against the same RM, we might also expect the gold score to peak at a substantially earlier KL distance, analogous to what we see when we scale the RM size (section 3.2), or for larger policies to more efficiently utilize the same number of RL feedback steps (section 3.3)[12].

One possible hypothesis is that, because RLHF can be viewed as Bayesian inference from the prior of the initial policy [Korbak et al., 2022][13], increases in policy size are only improving the modelling accuracy of the human demonstration distribution.

[11] We can think of noise as a particular case of this where the independent noise is correlated with signal+noise, but of course there is no causal relation between signal and noise.

[12] It is also not the case that the 6B policy run has higher KL distance for the same number of RL steps; in fact, we observe that it has *lower* KL distance for the same number of steps (fig. 15)

[13] The result of Korbak et al. [2022] concerns varying KL penalties rather than KL distances with no KL penalty, but as we observe in section 3.6, this is equivalent on our setting.

4.5 Limitations and Future Work

In addition to the overoptimization studied in this paper (due to the mismatch between the reward model and the ground truth labels), there exists another source of overoptimization due to mismatch between the ground truth labels and the actual human intent. This contains issues ranging from the mundane, such as labellers choosing options that only *appear* to match their intent[14], to substantially more philosophically fraught issues [Armstrong and Mindermann, 2018, Sunstein et al., 2001]. The main limitation of this work is that this additional source of overoptimization is not captured in the setting of this paper. See section 5 for discussion of related work in alignment.

Some additional limitations and future directions include:

- **Validating these results on other environments and experimental setups.** While the experiments in this paper all use the InstructGPT environment, the main value of these results lies in the extent to which they reflect general phenomema. Confirming whether these results generalize to other settings would be extremely valuable to that end.[15]
- **Validating the synthetic setting.** The synthetic setting might not transfer to real world settings, for instance because there is substantial correlation between RMs.
- **Investigating methods for making RMs more robust to optimization.** While there has been prior work in this direction (see section 5), there is still much work to be done in systematically investigating ways to make RMs more robust.
- **Exploring other forms of optimization and categorizing their differences.** While this work focuses exclusively on BoN and RL there are other ways of applying optimization pressure against a model of a reward signal, either implicit or explicit. This includes GeDi-like steering, Decision Transformers[16], variants of BoN like beam search, and other RL algorithms.
- **Better understanding the functional form of proxy RM scores.** In our modeling, we find that the proxy RM scores are more difficult to predict for both BoN and RL (section 3.2). While they seem to have a major linear component, there is sufficient variation that fitting a linear regression is not very good at predicting extrapolated proxy RM scores.
- **Exploring adversarial Goodhart empirically.** In this work we deal with systems not powerful enough to cause adversarial Goodhart. However, it is plausible that adversarial Goodhart is especially important, or is associated with phase changes that break the trends seen in this paper.
- **Exploring scaling with policy size in more detail.** Our exploration of policy size scaling in this paper was limited to only two policy sizes. It is possible that there exist trends not seen in our exploration when considering the policy size more carefully.
- **Exploring multi-iteration RLHF.** In particular, checking for deviations from the assumptions of section 4.3.

We hope this paper leads to future work further bridging conceptual and empirical alignment research.

5 Related Work

Goodhart's Law in its modern formulation was first introduced in Hoskin [1996], with many of the key ideas introduced in prior works [Campbell, 1969, Goodhart, 1975]. Many approaches have been proposed for reducing overoptimization in general [Taylor, 2016, Everitt et al., 2017], as well as in RMs [Gleave and Irving, 2022], including within the field of adversarial robustness [Chakraborty et al., 2018]. Overoptimization of reward models can be viewed as a special case of

[14] For instance, the example of a robotic hand learning from human feedback to only *appear* to grasp a ball, presented in `https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/` [Christiano et al., 2017]

[15] In the course of our experiments, we observed visually similar results on the WebGPT environment [Nakano et al., 2021].

[16] One could consider measuring the actual achieved ground truth/gold score achieved for each "proxy" score conditioned on, a la fig. 8, as testing the implicit reward-behavior mapping encoded by the model.

specification gaming (also known as reward hacking). Previous work has shown numerous examples of such behavior in a wide variety of settings [Krakovna et al., 2020, Lehman et al., 2020]. Pan et al. [2022] explores a diverse set of RL environments and finds phase transitions in some settings. A number of works have proposed theoretical models of Goodhart's Law and reward hacking [Krakovna and Kumar, 2019, Manheim and Garrabrant, 2018, Skalse et al., 2022], including Zhuang and Hadfield-Menell [2020] which exhibits very similar overoptimization curves as observed in this paper in some toy environments.

One can think of overfitting as a special case of Goodhart's law where the proxy is the score on some finite set of samples, whereas our actual objective includes its generalization properties as well. Overfitting has been observed and studied in RL settings [Zhang et al., 2018a,b, Farebrother et al., 2018, Cobbe et al., 2019]. Song et al. [2019] studies "observational overfitting" in RL settings, which is closely related to causal Goodhart [Manheim and Garrabrant, 2018].

Adversarial attacks and robustness are also very closely related fields. Many works have demonstrated the existence of adversarial examples in all kinds of neural networks [Szegedy et al., 2013, Lin et al., 2017, Ebrahimi et al., 2018, Dai et al., 2018], and proposed methods to measure and increase neural network robustness [Gu and Rigazio, 2014, Zheng et al., 2016, Carlini et al., 2019, Guo et al., 2021].

Scaling laws have seen substantial success in machine learning for predicting properties of language models [Kaplan et al., 2020, Henighan et al., 2020, Hernandez et al., 2021] and has led to better theoretical understanding of language models [Sharma and Kaplan, 2020, Bahri et al., 2021].

Reinforcement learning from human feedback [Christiano et al., 2017, Ibarz et al., 2018] has been used broadly in language models [Stiennon et al., 2020, Ouyang et al., 2022, Nakano et al., 2021, Bai et al., 2022]. It is also a first step towards recursive reward modelling [Leike et al., 2018], an approach towards reducing the additional source of overoptimization described in section 4.5, though it is subject to some theoretical limitations [Christiano et al., 2021]. We observe similar approximately-linear proxy RM scores observed in Bai et al. [2022][17], though we observe an early-KL bend in the proxy RM scores, and there are some occasional outliers with very small RMs and data sizes.

More broadly, AI alignment is the problem of ensuring that the goals of AI systems are aligned with the goals of humans [Ngo, 2022], including future AI systems which may exceed humans [Bostrom, 2014]. There are a number of reasons to expect AI misalignment, especially in those more powerful future systems, to occur [Omohundro, 2008, Turner et al., 2021, Armstrong et al., 2013, Hubinger et al., 2019, Soares et al., 2015], and to result in catastrophic outcomes [Carlsmith, 2022, Cotra, 2022].

Acknowlegements

We thank Vivek Hebbar, Jared Kaplan, Jan Leike, Kyle McDonell, Dan Mossing, Ethan Perez, Laria Reynolds, and Jeff Wu for valuable discussion and feedback.

References

Stuart Armstrong and Sören Mindermann. Occam's razor is insufficient to infer the preferences of irrational agents. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

Stuart Armstrong et al. General purpose intelligence: arguing the orthogonality thesis. *Analysis and Metaphysics*, 12(68):1–20, 2013.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[17] Note that Bai et al. [2022] scaled the policy size with the RM size, while we hold the policy size constant.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Inc., USA, 1st edition, 2014. ISBN 0199678111.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Donald T Campbell. Reforms as experiments. *American psychologist*, 24(4):409, 1969.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness, 2019. URL `https://arxiv.org/abs/1902.06705`.

Joseph Carlsmith. Is power-seeking AI an existential risk? *arXiv preprint arXiv:2206.13353*, 2022.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

Paul Christiano, Ajeya Cotra, and Mark Xu. Eliciting latent knowledge: How to tell if your eyes deceive you, 12 2021. URL `https://docs.google.com/document/d/1WwsnJQstPq91_Yh-Ch2XRL8H_EpsnjrC1dwZXR37PC8`.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1282–1289. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/cobbe19a.html`.

Ajeya Cotra. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover, 2022. URL `https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to`.

Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. Adversarial attack on graph structured data. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1115–1124. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/dai18b.html`.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. On adversarial examples for character-level neural machine translation. *arXiv preprint arXiv:1806.09030*, 2018.

Tom Everitt, Victoria Krakovna, Laurent Orseau, Marcus Hutter, and Shane Legg. Reinforcement learning with a corrupted reward channel. *arXiv preprint arXiv:1705.08417*, 2017.

Jesse Farebrother, Marlos C Machado, and Michael Bowling. Generalization and regularization in dqn. *arXiv preprint arXiv:1810.00123*, 2018.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements. 2022. URL `https://storage.googleapis.com/deepmind-media/DeepMind.com/Authors-Notes/sparrow/sparrow-final.pdf`.

Adam Gleave and Geoffrey Irving. Uncertainty estimation for language reward models. *arXiv preprint arXiv:2203.07472*, 2022.

Charles Goodhart. Problems of monetary management: the uk experience in papers in monetary economics. *Monetary Economics*, 1, 1975.

Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. Gradient-based adversarial attacks against text transformers, 2021. URL `https://arxiv.org/abs/2104.13733`.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Keith Hoskin. The "awful idea of accountability" : inscribing people into the measurement of objects. *Accountability : power, ethos and the technologies of managing / edited by Rolland Munro and Jan Mouritsen*, 1996.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31, 2018.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Tomasz Korbak, Ethan Perez, and Christopher L Buckley. Rl with kl penalties is better viewed as bayesian inference. *arXiv preprint arXiv:2205.11275*, 2022.

Victoria Krakovna and Ramana Kumar. Classifying specification problems as variants of goodhart's law, 8 2019. URL `https://vkrakovna.wordpress.com/2019/08/19/classifying-specification-problems-as-variants-of-goodharts-law/`.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of AI ingenuity, 4 2020. URL `https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity`.

Joel Lehman, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J. Bentley, Samuel Bernard, Guillaume Beslon, David M. Bryson, Patryk Chrabaszcz, Nick Cheney, Antoine Cully, Stephane Doncieux, Fred C. Dyer, Kai Olav Ellefsen, Robert Feldt, Stephan Fischer, Stephanie Forrest, Antoine Frénoy, Christian Gagné, Leni Le Goff, Laura M. Grabowski, Babak Hodjat, Frank Hutter, Laurent Keller, Carole Knibbe, Peter Krcah, Richard E. Lenski, Hod Lipson, Robert MacCurdy, Carlos Maestre, Risto Miikkulainen, Sara Mitri, David E. Moriarty, Jean-Baptiste Mouret, Anh Nguyen, Charles Ofria, Marc Parizeau, David Parsons, Robert T. Pennock, William F. Punch, Thomas S. Ray, Marc Schoenauer, Eric Shulte, Karl Sims, Kenneth O. Stanley, François Taddei, Danesh Tarapore, Simon Thibault, Westley Weimer, Richard Watson, and Jason Yosinski. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial life*, 26(2):274–306, 2020.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of adversarial attack on deep reinforcement learning agents, 2017. URL https://arxiv.org/abs/1703.06748.

David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Richard Ngo. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

Stephen M. Omohundro. The basic ai drives. In *Proceedings of the First Conference on Artificial General Intelligence*, pages 483–492. IOS Press, 2008. URL http://selfawaresystems.files.wordpress.com/2008/01/ai_drives_final.pdf.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. version 1.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*, 2020.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking, 2022. URL https://arxiv.org/abs/2209.13085.

Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Xingyou Song, Yiding Jiang, Stephen Tu, Yilun Du, and Behnam Neyshabur. Observational overfitting in reinforcement learning. *arXiv preprint arXiv:1912.02975*, 2019.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. *Computing Research Repository*, 2020. version 3.

Cass R Sunstein, Daniel Kahneman, David Schkade, and Ilana Ritov. Predictably incoherent judgments. *Stan. L. Rev.*, 54:1153, 2001.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Jessica Taylor. Quantilizers: A safer alternative to maximizers for limited optimization. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Alex Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23063–23074. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/c26820b8a4c1b3c2aa868d6d57e14a79-Paper.pdf.

Amy Zhang, Nicolas Ballas, and Joelle Pineau. A dissection of overfitting and generalization in continuous reinforcement learning. *arXiv preprint arXiv:1806.07937*, 2018a.

Chiyuan Zhang, Oriol Vinyals, Remi Munos, and Samy Bengio. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893*, 2018b.

Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 4480–4488, 2016.

Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned AI. *Advances in Neural Information Processing Systems*, 33:15763–15773, 2020.

A Proof of Regressional Goodhart identity

Lemma. *Let X and Z be independent absolutely continuous random variables with X normally distributed and either (a) Z normally distributed or (b) $|Z - \mathbb{E}[Z]| < \delta$ for some $\delta > 0$. Then for any real number c and as $\delta \to 0$,*

$$\mathbb{E}[X \mid X + Z = c] = \mathbb{E}[X] + (c - \mathbb{E}[X] - \mathbb{E}[Z]) \frac{\operatorname{Var}(X)}{\operatorname{Var}(X) + \operatorname{Var}(Z)} + \varepsilon,$$

where $\varepsilon = 0$ in case (a) and $\varepsilon = o(\operatorname{Var}(Z))$ in case (b).

Proof. First note that by making the substitutions $X' = X - \mathbb{E}[X]$ and $Z' = Z - \mathbb{E}[Z]$, we may assume without loss of generality that $\mathbb{E}[X] = \mathbb{E}[Z] = 0$. Let $\operatorname{Var}(X) = \sigma^2$ and $\operatorname{Var}(Z) = \tau^2$.

In case (a), the pair $(X, X + Z)$ is bivariate normal with covariance matrix

$$\begin{pmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 + \tau^2 \end{pmatrix},$$

and the result follows by standard properties of conditional distributions of multivariate normal distributions.

In case (b), let f_X and f_Z be the probability density functions of X and Z respectively. Then

$$\begin{aligned}
\mathbb{E}[X \mid X + Z = c] &= \frac{\int_{-\infty}^{\infty} (c - z) f_X(c - z) f_Z(z) \, \mathrm{d}z}{\int_{-\infty}^{\infty} f_X(c - z) f_Z(z) \, \mathrm{d}z} \\
&= c - \frac{\int_{-\delta}^{\delta} z \left(f_X(c) - f_X'(c) z + o(z)\right) f_Z(z) \, \mathrm{d}z}{\int_{-\delta}^{\delta} \left(f_X(c) - f_X'(c) z + o(z)\right) f_Z(z) \, \mathrm{d}z} \\
&= c - \frac{f_X(c) \mathbb{E}[Z] - f_X'(c) \mathbb{E}\left[Z^2\right] + o\left(\mathbb{E}\left[Z^2\right]\right)}{f_X(c) - f_X'(c) \mathbb{E}[Z] + o(1)} \\
&= c + \frac{f_X'(c)}{f_X(c)} \tau^2 + o\left(\tau^2\right) \\
&= c \left(1 - \frac{\tau^2}{\sigma^2}\right) + o\left(\tau^2\right) \\
&= c \left(\frac{\sigma^2}{\sigma^2 + \tau^2}\right) + o\left(\tau^2\right),
\end{aligned}$$

as required. □

B RL form details

Ideally all overoptimization forms would have finite slope at the origin. We tried the following forms:

- $d\left(\alpha_{\mathrm{RL}} - \beta_{\mathrm{RL}} \log\left(1 + d\right)\right)$: Has slope α at the origin; however, has substantially worse extrapolation behavior. We can replace the 1 with a learned ϵ but that introduces another degree of freedom.
- Power laws $d\left(\alpha_{\mathrm{RL}} - \beta_{\mathrm{RL}} d^{\gamma_{\mathrm{RL}}}\right)$: Has slope α at the origin; however, this adds another degree of freedom, and the best fits resulted in small values of γ_{RL}.

Note that the power law forms with small γ_{RL} approximate the RL form that we decided on, as $\lim_{n\to\infty} n(x^{1/n} - 1) = \log x$.

C Hyperparameters

Hyperparameter	Value
RM Adam learning rate multiplier	1.67e-2
RM batch size	64
RL Adam learning rate multiplier	4e-3
RL batch size	256
RL PPO clipping parameter	0.2
RL Timesteps per rollout	256
RL minibatches per epoch	128
RL GAE bootstrapping parameter	0.95

Table 1: Hyperparameters used throughout the experiments.

What is full of holes but still holds water?

n	Generation	Proxy score	Gold score
1	Crime\nMussels are typically found to be full of holes but will the moisture still be retained and will it hurt my eyes? The general response is yes as the clam is generally of a moistest nature. However, it all depends on the type of clam you are purchasing.<|endoftext|>	-0.1922	-0.5225
3	Most likely a pipe is having trouble staying full.<|endoftext|>	0.0322	-0.0165
10	A sponge<|endoftext|>	0.2336	0.4828
30	When something is full of holes, it is used for stirring or moving liquid.<|endoftext|>	0.6534	-0.1543
100	A tornado is usually a swirling cloud of swirling air with a spinning jet which stirs up winds full of sharp debris.<|endoftext|>	0.8968	-0.3367
300	A tornado is usually a swirling cloud of swirling air with a spinning jet which stirs up winds full of sharp debris.<|endoftext|>	0.8968	-0.3367
1000	A tornado is usually a swirling cloud of swirling air with a spinning jet which stirs up winds full of sharp debris.<|endoftext|>	0.8968	-0.3367
3000	A bore hole is a hole drilled into a rock for the purpose of exploring a fossil-bearing sedimentary or bedrock deposit.<|endoftext|>	0.9003	0.2733
10000	A bore hole is a hole drilled into a rock for the purpose of exploring a fossil-bearing sedimentary or bedrock deposit.<|endoftext|>	0.9003	0.2733
30000	A pothole is a structural vulnerability that allows water to penetrate its cavity and cause damage to passing vehicles or the surface it rests on.<|endoftext|>	0.9527	0.5490

Table 2: A sample of the BoN answers on a single InstructGPT question (policy=1.2B, proxy RM=12M). For each individual question, the gold scores do not follow as clean a trend as they do when averaged over many questions as in fig. 1.

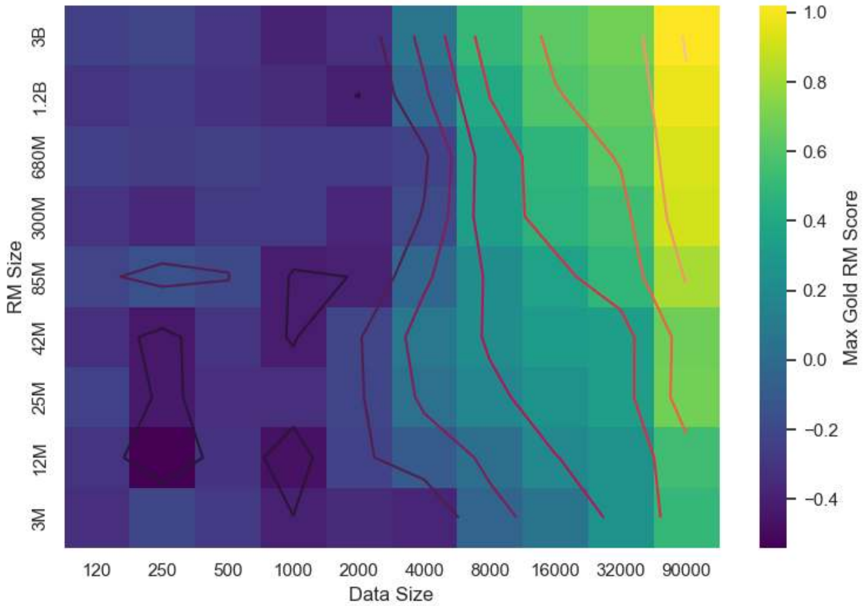


Figure 10: Maximum gold scores for all RM size and data size combinations.

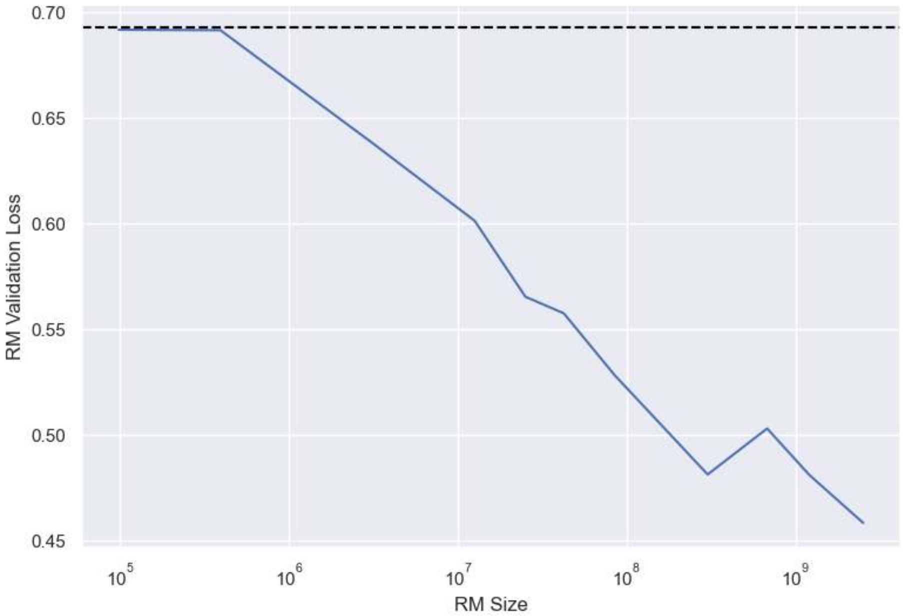


Figure 11: Validation losses for the proxy RMs in section 3.2 by size, plus the two near-chance level RMs.

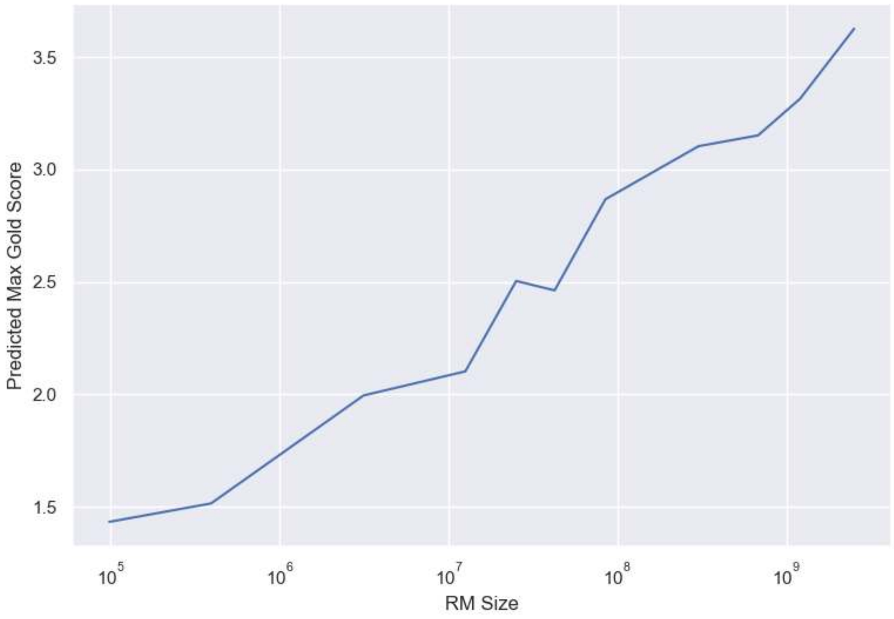


Figure 12: Max BoN gold scores ($\alpha_{\text{bon}}/2\beta_{\text{bon}}$) predicted with the BoN closed form

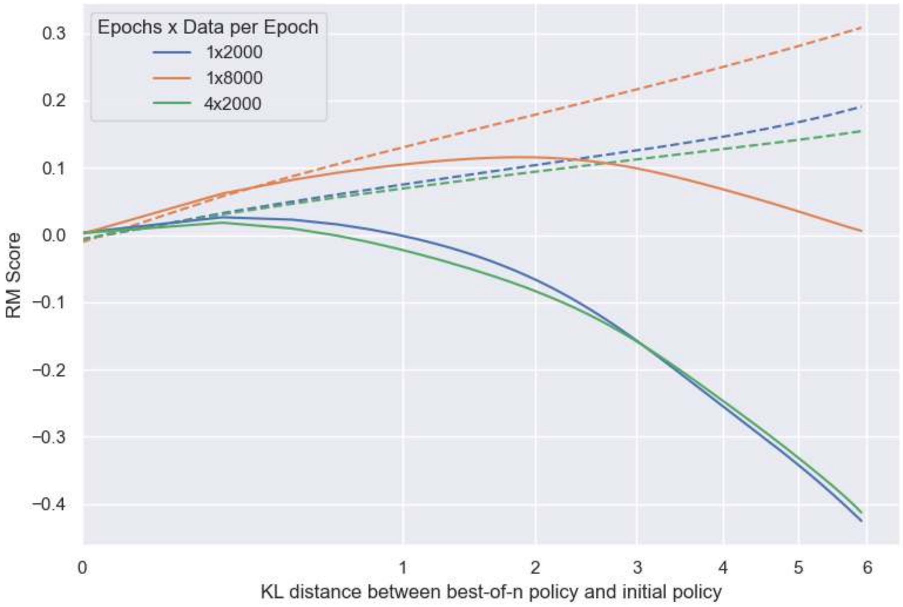


Figure 13: Total number of data points seen does not seem to affect the gold RM score much compared to the number of unique data points seen. Averaged across RM sizes. The numbers of datapoints (2000–8000) is intentionally chosen to straddle the sharp increase in performance. The validation loss of the 1x2000, 1x8000, and 4x2000 RMs are 0.686109, 0.654857, and 0.683869 respectively.

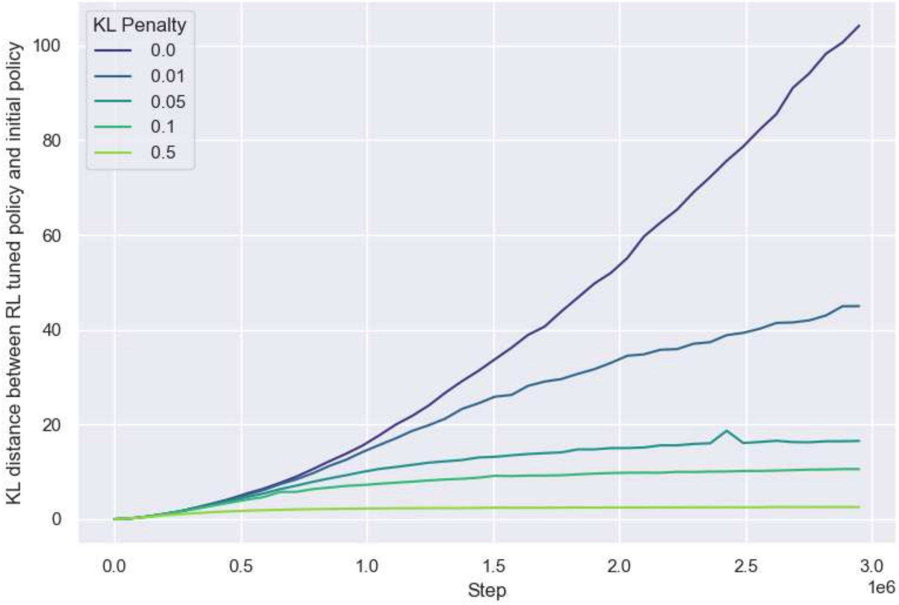


Figure 14: Change in $\mathrm{KL}_{\mathrm{RL}}$ throughout RL training for various different KL penalties. We observe that KL distance increases approximately monotonically with step count, and converges for higher KL penalties.

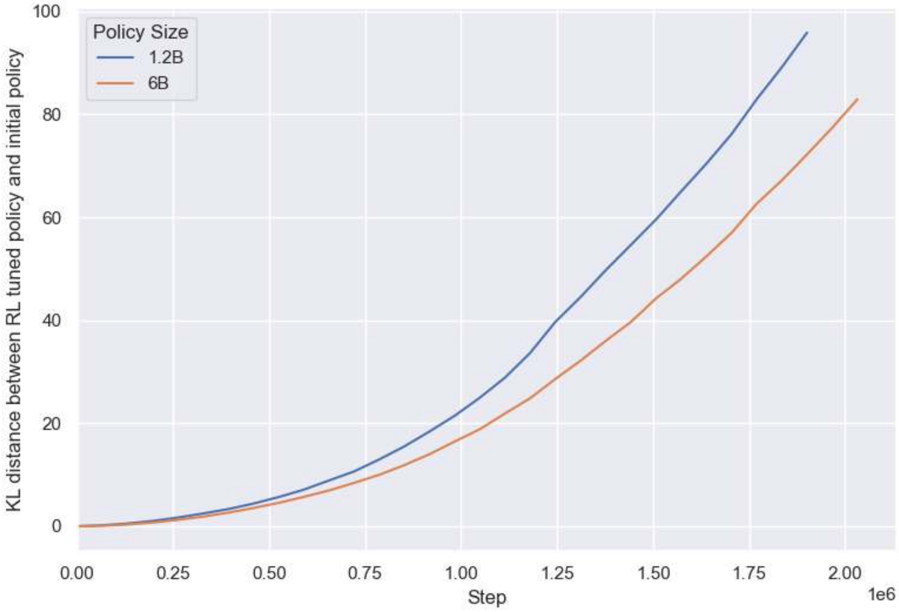


Figure 15: $\mathrm{KL}_{\mathrm{RL}}$ with policy size (RM size = 12M)

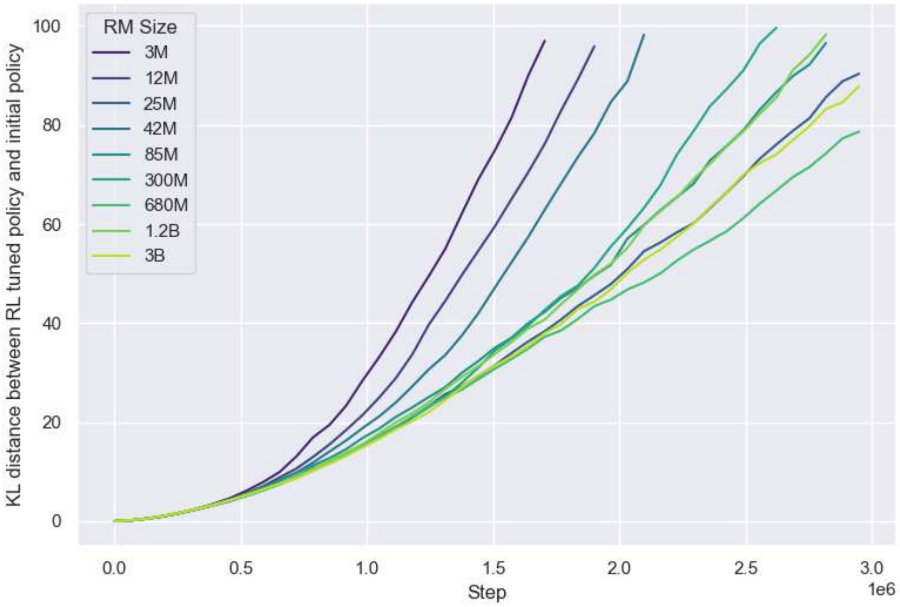


Figure 16: $\mathrm{KL}_{\mathrm{RL}}$ with RM size

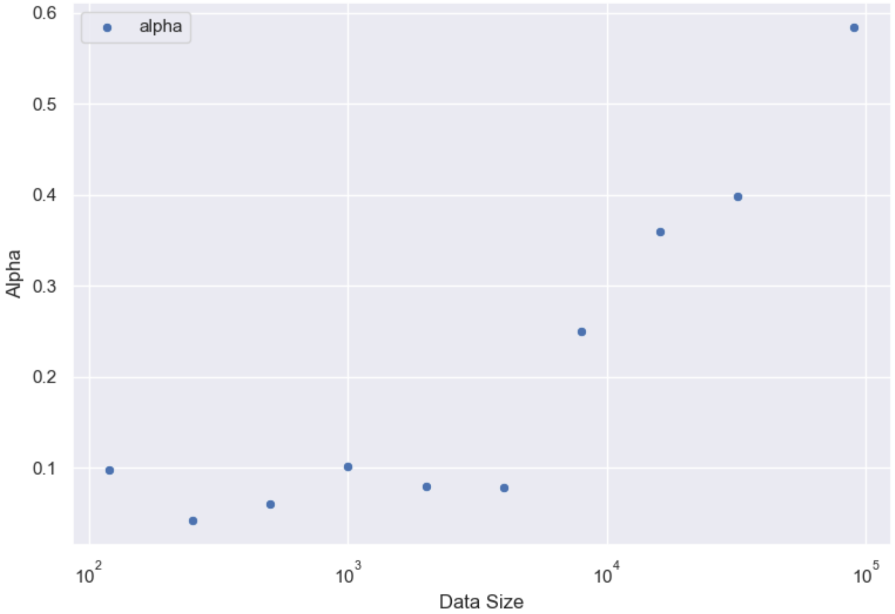


Figure 17: α_{bon} with dataset size, averaged across RM sizes

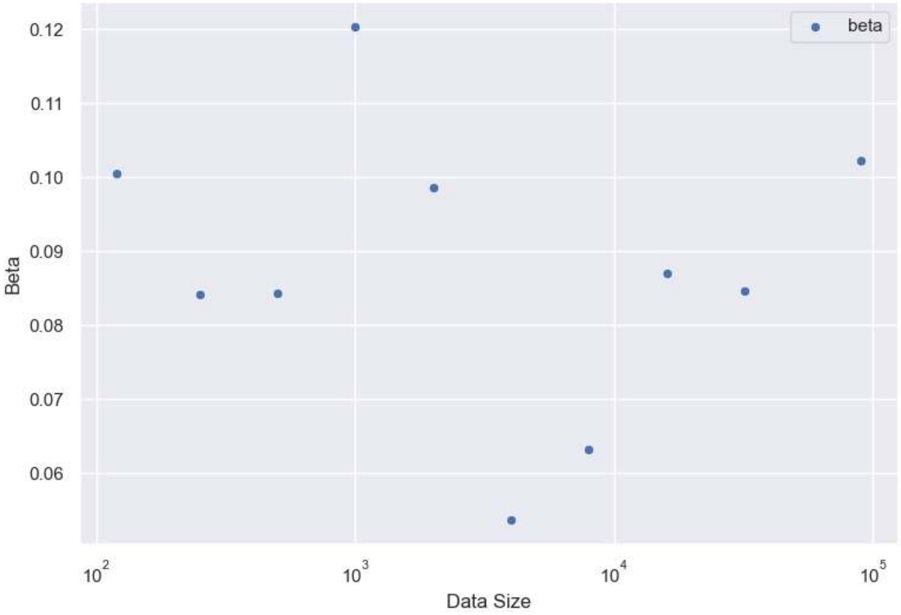


Figure 18: $\beta_{\mathrm{bo}n}$ with dataset size, averaged across RM sizes

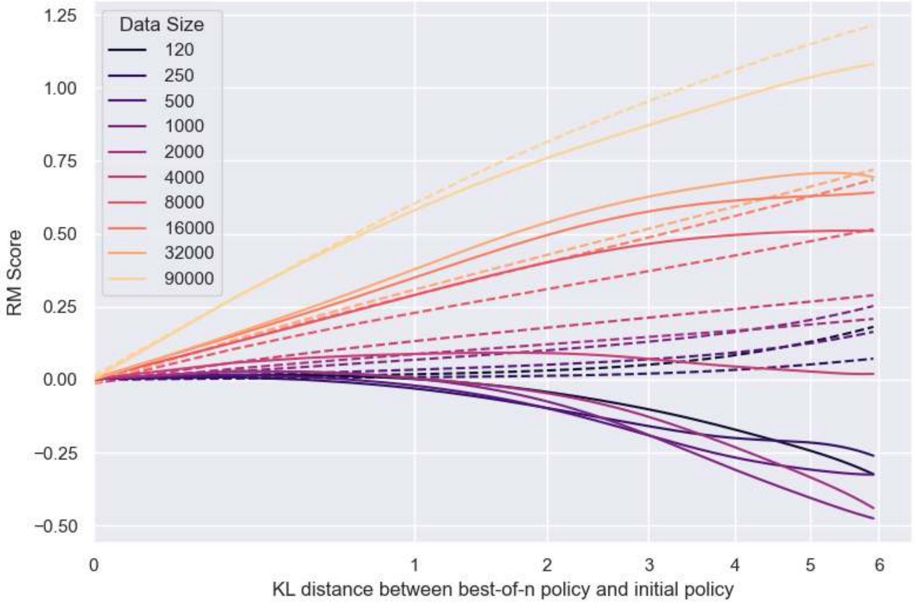


Figure 19: RM data scaling experiments, BoN, RM size=3B

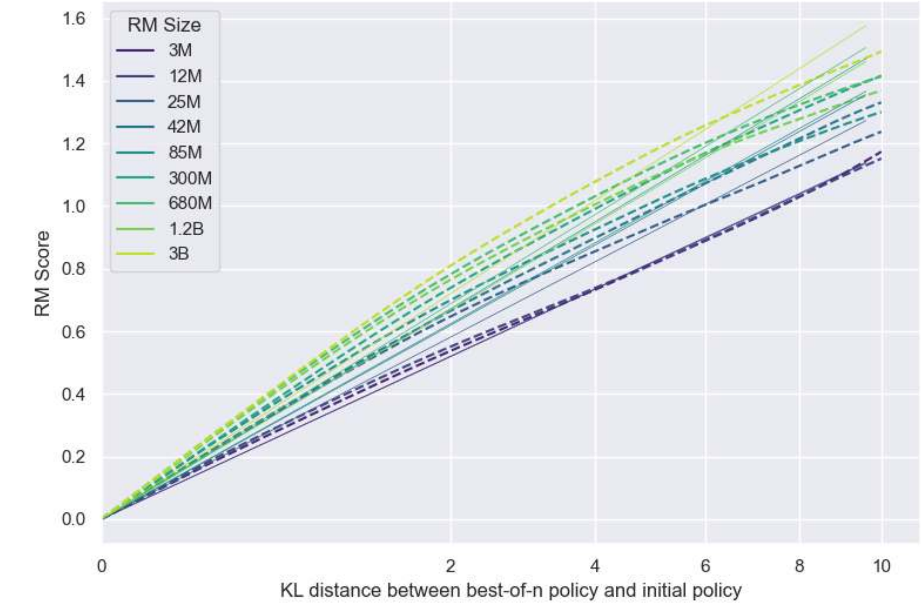


Figure 20: The BoN proxy scores are slightly concave, so that a linear fit does not fit well.

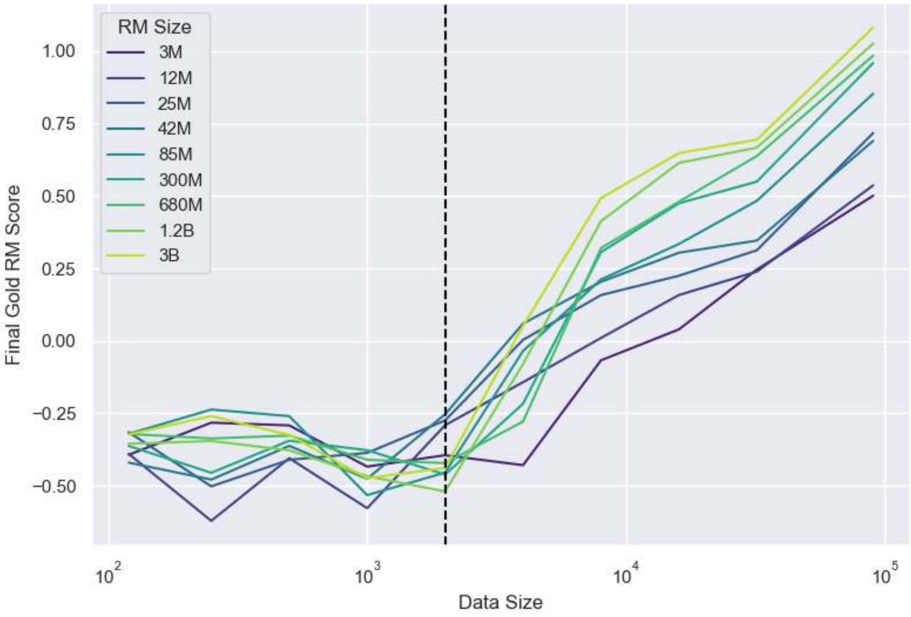


Figure 21: BoN Gold scores at n=1,000, broken down by data size and RM size. See fig. 6 for RM losses. Vertical dotted line approximately indicates first better-than-random data size.

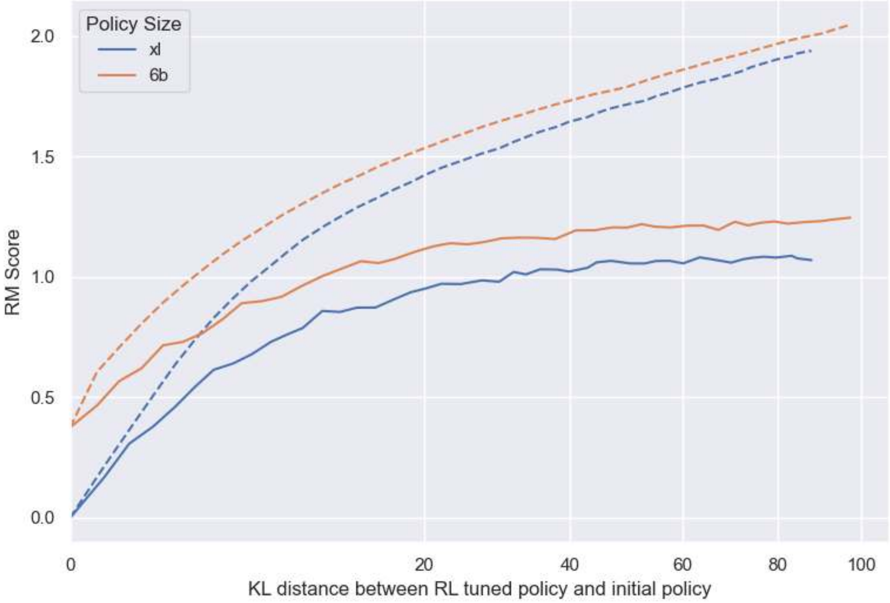


Figure 22: RL experiments with 3B RM and different policy sizes.

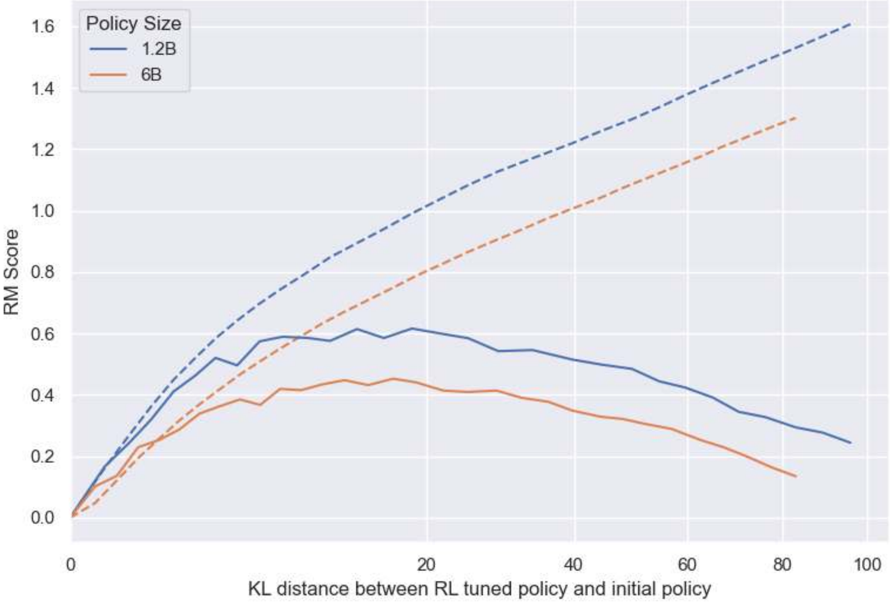


Figure 23: fig. 7b with all runs normalized from 0.

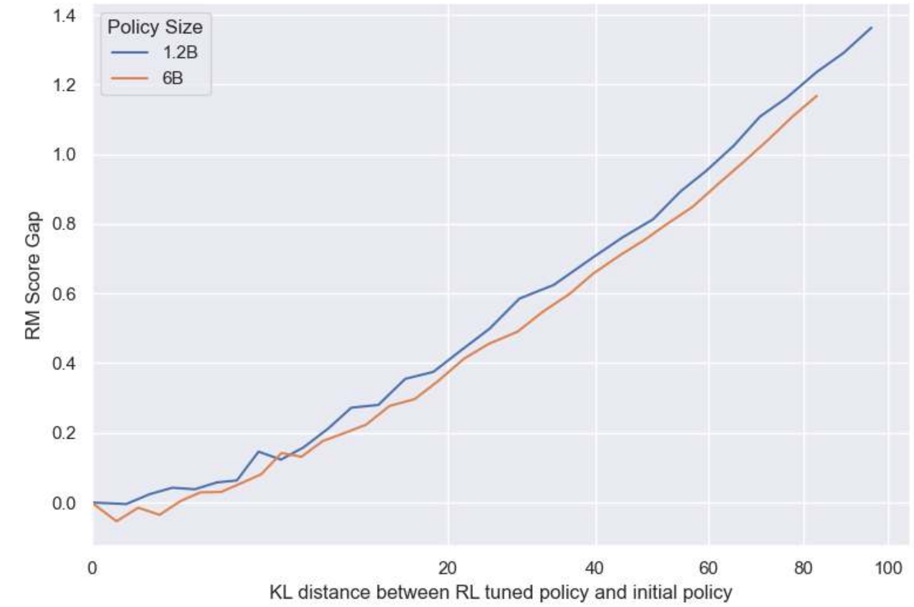


Figure 24: The gap between the proxy and gold scores in the RL policy sweep (fig. 24).

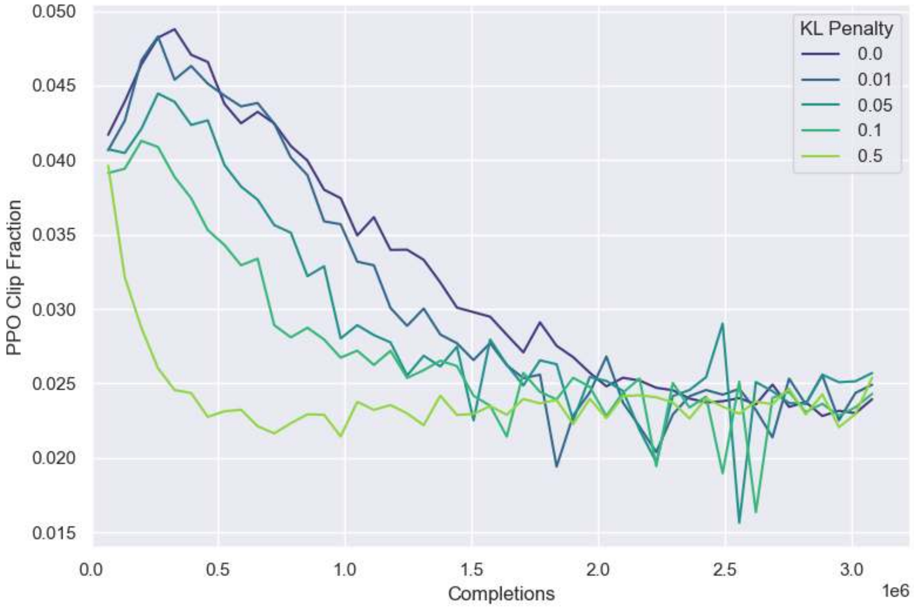
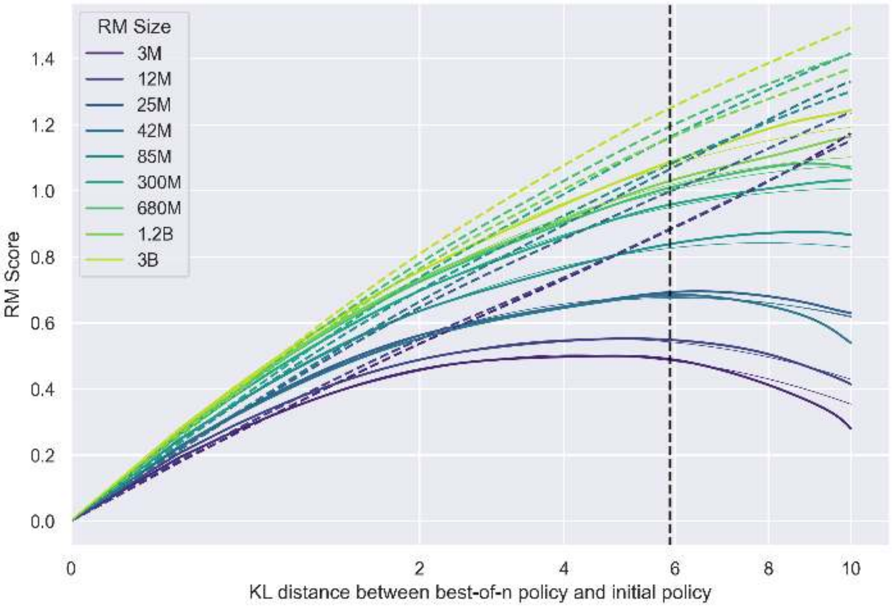


Figure 25: The fraction of updates clipped by PPO.



(a) BoN

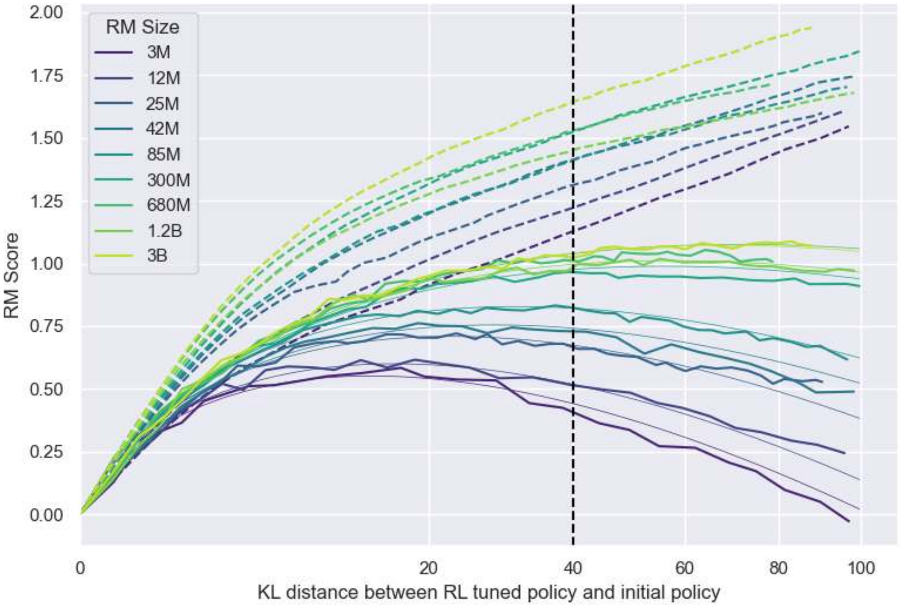


Figure 26: Extrapolation quality of fits in fig. 1. The regressions (shown in faint lines) are only fit to data to the left of the vertical black dotted lines. In the case of BoN, this represents a true advance prediction, as the functional form was chosen without collecting any data past a KL of 6 nats.