Countering Reward Over-optimization in LLM with Demonstration-Guided Reinforcement Learning

Mathieu Rita[*]
INRIA, Paris

Florian Strub
Cohere

Rahma Chaabouni
DeepMind

Paul Michel
DeepMind

Emmanuel Dupoux
EHESS,ENS-PSL,CNRS,INRIA
Meta AI Research

Olivier Pietquin
Cohere

Abstract

While reinforcement learning (RL) has been proven essential for tuning large language models (LLMs), it can lead to reward over-optimization (ROO). Existing approaches address ROO by adding KL regularization, requiring computationally expensive hyperparameter tuning. Additionally, KL regularization focuses solely on regularizing the language policy, neglecting a potential source of regularization: the reward function itself. Inspired by demonstration-guided RL, we here introduce the Reward Calibration from Demonstration (RCfD), which leverages human demonstrations and a reward model to recalibrate the reward objective. Formally, given a prompt, the RCfD objective minimizes the distance between the demonstrations' and LLM's rewards rather than directly maximizing the reward function. This objective shift avoids incentivizing the LLM to exploit the reward model and promotes more natural and diverse language generation. We show the effectiveness of RCfD on three language tasks, which achieves comparable performance to carefully tuned baselines while mitigating ROO.

1 Introduction

Reinforcement learning (RL) has long been used to train conversational agents, ranging from designing dialogue strategies (Singh et al., 1999; Lemon and Pietquin, 2007) to language modelling (Ouyang et al., 2022). While supervised learning excels at pre-training LLMs (Achiam et al., 2023; Touvron et al., 2023), RL stands out for finetuning LLMs. It allows optimizing non-differentiable objectives (Ranzato et al., 2016; Paulus et al., 2018), improving sequence-planning in goal-oriented dialogues (Wei et al., 2018; Strub et al., 2017), or aligning LLMs with human preferences (RLHF) (Christiano et al., 2017; Ouyang et al., 2022), which leads to more helpful and harmless LLMs (OpenAI, 2023; Bai et al., 2022).

Finetuning LLMs with RL typically involves scoring their utterances with a reward function, which is then maximized using online RL methods. Unfortunately, this optimization process is known to be brittle if not carefully controlled (Lewis et al., 2017), reducing language diversity (Gao et al., 2023), generating unnatural language patterns to artificially inflate rewards (Paulus et al., 2018), or altering the LLM semantics and syntax (Lazaridou et al., 2020). This phenomenon has recently been referred to as reward over-optimization (ROO)[1].

ROO may be mitigated by incorporating a KL-regularization term to anchor the finetuned model to its initial human-like language policy (Christiano et al., 2017). However, calibrating the KL term requires careful hyperparameter tuning, which is computationally expensive when finetuning LLM with online RL (Stiennon et al., 2020). Offline approaches, such as Direct Preference Optimization (DPO) (Rafailov et al., 2023), attempt to address ROO by bypassing the reward estimation and directly maximizing user preferences through pairwise comparisons. Unfortunately, these methods also suffer from ROO, albeit not optimizing the reward explicitly, and still requires careful KL regularization (Azar et al., 2023; Tunstall et al., 2023). In other words, these attempts to address ROO primarily focus on constraining the language policy, leaving the reward objective itself unaddressed. Besides, it cannot be generalized beyond pairwise data, limiting it to RLHF settings only.

This paper proposes a novel approach, Reward Calibration from Demonstration (RCfD), to tackle ROO in LLMs. Inspired by demonstration-guided RL (Schaal, 1996; Pertsch et al., 2021), RCfD utilizes human demonstrations and a reward model

---

[*] Corresponding Authors: mathieu.rita@inria.fr, florian@cohere.com

[1] ROO may englobe various language optimization artifacts such as reward hacking (Skalse et al., 2022), language drift (Lu et al., 2020) or overfitting (Zhang et al., 2018).

Findings of the Association for Computational Linguistics ACL 2024, pages 12447–12472
August 11-16, 2024 ©2024 Association for Computational Linguistics

to guide the LLM towards generating outputs that achieve similar rewards to those of the demonstrations. This shift from directly maximizing the reward function to calibrating it based on demonstrations helps prevent LLMs from exploiting the reward model and encourages more natural language generation. Furthermore, unlike pure imitation learning, RCfD operates at the sequence level, mitigating exposure bias (Ranzato et al., 2016) and promoting greater diversity in the generated text.

We conducted a series of experiments to investigate the effectiveness of RCfD. First, we apply RCfD to maximize the language model's sequence log-likelihood with RL. This experiment demonstrates that RCfD prevents the language degeneration typically observed in RL while avoiding the compounding errors associated with imitation learning. Next, we optimize RCfD objectives on two RL language tasks, achieving performance comparable to tuned baselines. This showcases RCfD's ability to effectively address ROO while maintaining task performance. Finally, we explore RCfD in multi-reward settings, where the goal is to optimize multiple, potentially conflicting rewards. By targeting a point on the Pareto frontier through demonstrations, RCfD controls the optimization process. Our experiments provide strong evidence that recalibrating the reward objective with demonstrations mitigates ROO and offers a promising approach for tackling complex language RL tasks where human demonstrations are available.

2 Related Works

Demonstration-Guided RL (DGRL) aims at interleaving expert data with a reward objective for sequence planning (Schaal, 1996; Ramírez et al., 2022). Unlike imitation learning, which directly copies expert actions, DGRL uses demonstrations as a guiding force to address common RL challenges. For instance, expert trajectories may guide exploration (Nair et al., 2018; Hester et al., 2018), help to discover high-level policy skills (Pertsch et al., 2021), or improve sample efficiency (Rajeswaran et al., 2018; Hester et al., 2018). DGRL has been used in robotics to prevent overfitting to a simulated environment and ensure realistic robot movements. For example, (Peng et al., 2018) added a reward term to limit the distance motion with the demonstrations, or (Zhu et al., 2018) included an extra discriminative reward to detect when the trajectory does not match the demonstration. Inspired by these successes, our proposed algorithm, RCfD, leverages DGRL to tune LLMs while mitigating the risk of ROO.

Reward Over-Optimization (ROO) RL was successfully used in multiple language tasks ranging from language modelling (de Masson d'Autume et al., 2019), translation (Ranzato et al., 2016; Bahdanau et al., 2017), summarization (Stiennon et al., 2020), code generation (Le et al., 2022), instruction following (Ouyang et al., 2022) or question answering (Nakano et al., 2021). However, RL methods were quickly reported to exploit language metrics (Wu and Hu, 2018), either creating emergent language (Lewis et al., 2017; Strub et al., 2017) or overfitting text classifiers (Ramamurthy et al., 2022) and user preference models (Gao et al., 2023). More generally, whenever maximizing the reward function over a certain point starts lower the ground truth performance, this can be referred to as reward over-optimization (ROO) (Gao et al., 2023; Moskovitz et al., 2023). ROO has two main origins: (i) the absence of grounding: as solely trained on optimizing scores, LLMs can become detached from human language (Lee et al., 2019; Lazaridou et al., 2020), (ii) the optimization of imperfect reward models (Schatzmann et al., 2006).

Countering ROO Reward over-optimization is often mitigated by tying the finetuned model to its base distribution through a KL regularization (Ziegler et al., 2019; Ouyang et al., 2022; Bai et al., 2022). While simple, this method has multiple variants, e.g., using the KL in the loss (Glaese et al., 2022) or in the reward (Roit et al., 2023), using decay heuristics (Ziegler et al., 2019) or altering the referent distributions (Noukhovitch et al., 2023). However, KL-tuning requires costly cross-validation as it is impossible to predict the final impact of the KL regularization before training (Ramé et al., 2024). As RCfD targets the reward distributions from demonstrations, the resulting LLM behavior is far more predictable, making it an *a-priori* regularization method as explored in 5.3.

Another strategy implemented by DPO (Rafailov et al., 2023) avoids modeling the reward function by leveraging pairwise comparisons. While this circumvents reward imperfection issues, DPO remains susceptible to overfitting (Tunstall et al., 2023), with KL-regularization only marginally regularizing the training (Azar et al., 2023). More generally, DPO is designed explicitly for optimizing rewards derived from preference models, mak-

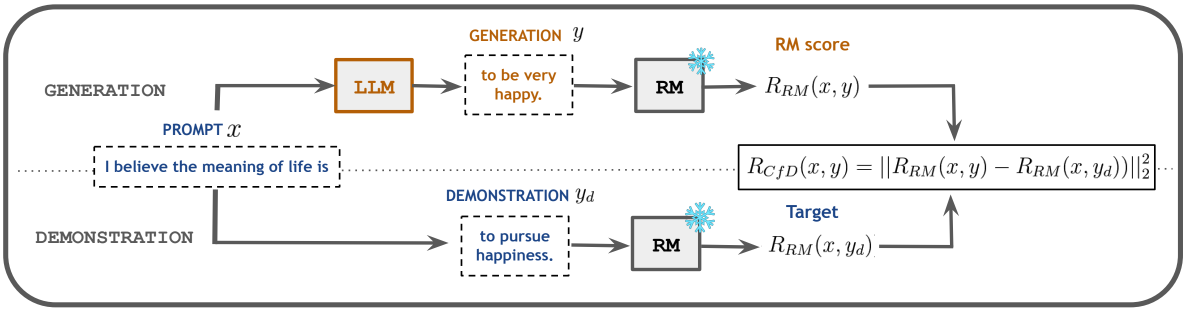


Figure 1: The RCfD objective is the L2-distance between the reward from the LM and the reward from the demonstration. Given a prompt x, a demonstration y_d, and the LLM continuation y, the RM computes the demonstration reward $R_{RM}(x, y_d)$, and the LM reward $R_{RM}(x, y)$. Instead of maximizing $R(x, y)$ as in standard RL, we here aim at maximizing the RCfD objective defined as $R_{RCfD}(x, y) = -||R_{RM}(x, y) - R_{RM}(x, y_d)||_2^2$.

ing it inoperable in RL language tasks where LLM completions are scored individually, e.g., success scores (Le et al., 2022) or classifiers (Roit et al., 2023). In particular, we explore in section 5.1 and 5.2 two settings where DPO cannot applied, demonstrating the interest of RCfD beyond the restricted use cases of RLHF.

Closer to our work, Moskovitz et al. (2023) identify proxy points where ROO occurs and retrain the LLM by dynamically reweighting the rewards not to exceed the proxy points, avoiding the ROO regime. While Moskovitz et al. (2023) and RCfD both recalibrate the reward, RCfD avoid computing the proxy points using demonstrations, requiring less compute and no gold-standard metrics.

3 Reward Calibration from Demonstration

3.1 Notations and background

RL for LLM: Given a prompt x, the LLM auto-regressively generates a sequence of tokens y following the policy $\pi_\theta(.|x)$, where π_θ is a parametrized probability distribution. The prompt and its completion are assessed by a reward model (RM) R_{RM}. In RL, our goal is to find the optimal policy π_{θ^*} that maximizes the average reward model score over a dataset of prompts $\mathcal{D}$:

$$\pi_{\theta^*} = \text{argmax}_\theta \, \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(.|x)} \left[R_{RM}(x, y)\right]. \tag{1}$$

A KL regularization term is often added on top of the reward to prevent the language agent from diverging too much from its initial distribution:

$$R_\beta(x, y) = R_{RM}(x, y) - \beta \log\left(\frac{\pi_\theta(y|x)}{\pi_{\theta_0}(y|x)}\right) \tag{2}$$

where β is a training hyperparameter and π_{θ_0} the policy before RL fine-tuning. We here maximize the RL objective using Proximal Policy Optimization (PPO) (Schulman et al., 2017).

DGRL for LLM: DGRL combines a demonstration dataset in addition to the reward function in RL. Formally, the dataset $\mathcal{D} = (x^n, y_d^n)_{n=1}^N$ is composed of N pairs of prompts x and demonstrations y_d. Given the RM and a prompt x, we note $R_{RM}(x, y)$ the reward of the LLM's completion, and $R_{RM}(x, y_d)$ the reward of the demonstration.

3.2 RCfD Objective

Based on the dataset $\mathcal{D}$, we introduce the Reward Calibration from Demonstration (RCfD) objective:

$$R_{CfD}(x, y) = -||R_{RM}(x, y) - R_{RM}(x, y_d)||_2^2. \tag{3}$$

We omit the dependence to y_d in the R_{CfD} objective for simplicity. The complete pipeline from data to reward is depicted in Figure 1. Finally, when dealing with composite rewards as in section 5.3, we independently recalibrate and whiten each reward before summing them, i.e., $R_{CfD}(x, y) = \sum_i \sigma\left(r_{CfD}^i(x, y)\right)$ where $\sigma(.)$ is a whitening transformation and $r_{CfD}^i(x, y)$ the calibrated rewards.

By maximizing $R_{CfD}(x, y)$ instead of directly maximizing $R_{RM}(x, y)$, the LLM is trained to generate outputs that achieve a score similar to the expert demonstrations y_d. Consequently, this approach inherently avoids excessive optimization of the reward model. Rather than aiming for the highest possible RM score, the language model is trained to seek RM scores comparable to those achieved by the provided demonstrations.

4 Experimental Setting

We first use the log-likelihood optimization problem in LLMs to closely examine the issues that arise with standard RL and imitation learning. This helps us better understand the motivation behind RCfD. Then, we evaluate RCfD in a single reward setting. We confirm that RCfD performs similarly to existing best baselines while mitigating ROO. Finally, we show that RCfD successfully handles multi-reward objectives by using demonstrations to guide LLMs toward the desired behavior.

4.1 Use case 1: Building Intuition by Calibrating sequence-level log-likelihood

Motivation: While LLMs are trained to maximize their per-token log-likelihood (Williams and Zipser, 1989), they must generate entire sequences of words during inference. This regime mismatch can lead the LLM to accumulate errors over long sequences (Bengio et al., 2015). This phenomenon, namely exposure bias, may be lessened by maximizing the sequence level likelihood (Ranzato et al., 2016). However, if the sequence likelihood is over-optimized, the LLM can become prone to language degenerescence (Holtzman et al., 2020). This is called the sequence likelihood calibration problem (Zhao et al., 2023), and we here see how RCfD solves this calibration issue.

Setup: We cast the sequence likelihood calibration problem as an RL problem. Given a text context x and its continuation y, we define the reward function as $R_{RM}(x, y) = -\frac{1}{|y|}\log(\pi_{\theta_0}(x|y))$ where π_{θ_0} is a frozen pretrained LLM, and $|y|$ the number of generated tokens. Hence, the resulting agent should generate sequences that maximize the sequence log-likelihood of the frozen model.

We use the Wikipedia dataset (Wikimedia, 2023) where each text segment is split into prompt-continuation pairs with respective lengths of 64 and maximum 320 tokens. We use the continuation as a demonstration y_d for the RCfD objective. The LLM is a LlaMa2-7B (Touvron et al., 2023). Notably, the policy may generate up to 320 tokens during training but is evaluated with generations of up to 1000 at evaluation time to show the log-likelihood discrepancy.

Experiments: We optimized the LLM with either the R_{RM} or the R_{CfD} objective using PPO. We also performed Supervised Finetuning (SFT) on top of the Wikipedia demonstration.

4.2 Use case 2: Mitigating ROO in single reward settings

Motivation: Finetuning a LLM against a pretrained reward model is prone to ROO (Ziegler et al., 2019). We assess our RCfD's ability to recalibrate the reward objective to mitigate ROO while having strong downstream performances.

Setup: We showcase RCfD with two reward model settings: classifier RM (1), RLHF RM (2).

For the classifier RM (1), we train the LLM to generate positive movie reviews as in (Ramamurthy et al., 2022). The prompts x are the first 10 tokens from a positive review in the IMDB dataset (Maas et al., 2011), and the remaining tokens act as demonstrations y_d. The dataset is divided into training and validation sets. The policy is a LlaMa2-7B (Touvron et al., 2023). The reward model is a DistilBERT (Sanh et al., 2019) pretrained for sentiment classification on movie reviews[2]. The reward $R_{RM}(x, y)$ is the RM's output logit corresponding to the positive class. The maximum generation length is 160 tokens.

For the RLHF RM (2), we investigate the summarisation task. We use the TL;DR Reddit dataset (Völske et al., 2017), where annotators have ranked two generated summaries. As in (Lee et al., 2023), we filter the dataset to include only samples with high annotator confidence (≥ 5). This results in a collection of $22k$ prompts paired with their the chosen summary demonstration y_d. The policy is an Alpaca LLM (Taori et al., 2023), a LlaMa7b finetuned on instructions. The reward model is OpenAssistant's DeBerta model (Köpf et al., 2023) trained on multiple human preference datasets, including the TL;DR Reddit (Völske et al., 2017). The reward $R_{RM}(x, y)$ is the score computed by the preference model when processing x and y.

Experiments: We optimize R_{CfD}, $R_{\beta=0}$, $R_{\beta=0.1}$, and R_{β^*} objectives with PPO, where β^* was found by cross-validation to match the reward distribution. We add SFT baseline training on the demonstrations. For the RLHF setting (2), we also add a DPO baseline.

4.3 Use case 3: Multi-reward calibration

Motivation: When scaling language tasks, the training objective may combine multiple reward models together, e.g., balancing helpfulness and harmfulness (Bai et al., 2022; Glaese et al., 2022). This joint optimization presents the challenges:

---

[2] https://huggingface.co/lvwerra/distilbert-imdb

(1) correctly weighting the importance of each reward and (2) avoiding individual reward overoptimization (Moskovitz et al., 2023; Rame et al., 2023). We here show that RCfD naturally tackles both of these challenges by aligning the policy reward distribution on the demonstrations.

Setup: To study the multi-reward setting, we extend the summarization task (see 4.2) with a sequence length objective. We introduce the sequence length reward $R_{length}(x, y) = -|y|$ where $|y|$ is the number of tokens in the completion y, to penalize long token generation. Thus, the reward function to optimize is $R_\alpha = R_{RM}(x, y) + \alpha R_{length}(x, y)$. One must tune α to best compromise between the number of tokens and the preference. As noted in the sec 3.2, the RCfD objective automatically recalibrates both rewards by using the demonstration and without tuning any α. Finally, we apply the same setting described in 4.2.

Experiments: We use the same baselines as 4.2. Since DPO does not include length regularization, we report the checkpoint nearest to the demonstrations in terms of both R_{RM} and R_{length}.

4.4 Training and Evaluation

During finetuning, we use Low Rank Adaptation (LoRA) (Hu et al., 2022) with PEFT (Mangrulkar et al., 2022), and Adam optimizer (Kingma and Ba, 2014). In each case, we report the best model with the highest average reward on the evaluation set $\mathcal{D}_{val}$ after performing a grid search over the learning rate, batch size, LoRA parameters and α when applicable. Hyperparameters are reported in Appendix A, and the code is available at `https://github.com/MathieuRita/llm_demonstration_guided_rl`. We evaluate the models over three sets of metrics:

Average Reward: It measures the average reward R_{RM} from the LM on the validation set.

Reward distribution alignment: It measures the alignment between the distribution of rewards obtained by the LM and the distribution of rewards of demonstrations over the validation set. Formally, given the normalized distribution ρ_{π_θ} of rewards obtained when generating the continuation of validation prompts with stochastic sampling, and the normalized distribution of rewards of the demonstrations ρ_d, we define the alignment score $\mathcal{A}$ as the KL divergence between the two distributions, i.e., $\mathcal{A} = D_{KL}(\rho_d || \rho_{\pi_\theta})$. The lower $\mathcal{A}$, the more ρ_{π_θ} and ρ_d are aligned, with an optimal score of 0.

Model-based evaluations: We evaluate several features of the generations with an AI feedback process conducted by chat-Llama-70B (Touvron et al., 2023) as a judge. For each assessed feature, we provide the judge with the prompt-completion pair and a scoring question. To assess movie review generation, we evaluate the *task success* (is the review positive?) and *naturalness* (how human-like is the review?). To assess summarization, we ask the judge to evaluate the summary's *success*, *factuality*, *naturalness*, and *verbosity*. We report detailed feature descriptions and judge prompting in Appendix B. Furthermore, we introduce Δ_{demo}, computing the sum of the absolute differences between the model-based evaluation scores of the evaluated model and those of the demonstrations.

5 Results

In this section, we derive the results of the three use cases: sequence-level log-likelihood calibration problem, single-reward optimization, and multi-reward optimization.

5.1 Use case 1: Building intuition by Calibrating sequence-level log-likelihood

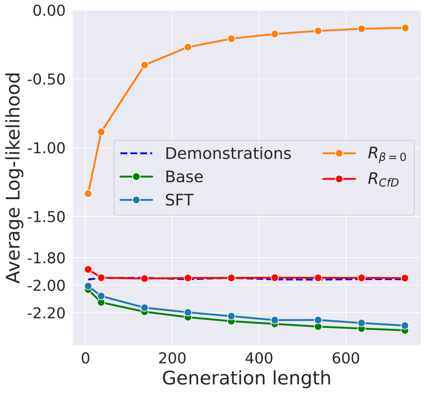


Figure 2: Average log-likelihood as a function of the generation length. Optimizing $R_{\beta=0}$ finds LLM exploits to minimize the likelihood, while imitation-based models suffer from exposure bias. Only R_{CfD} has an average log-likelihood that matches human behavior.

Sequence Log-likelihood lessening with SFT In Figure 2, the average log-likelihood of sentences generated by the initial LLM diminishes with longer sentences (green line). On the contrary, when we evaluate the log-likelihood of the demonstrations with the initial model, we do not observe this log-likelihood loss along the generation. This shows the existence of the exposure bias. Importantly, this exposure bias is barely reduced when performing SFT with the demonstra-

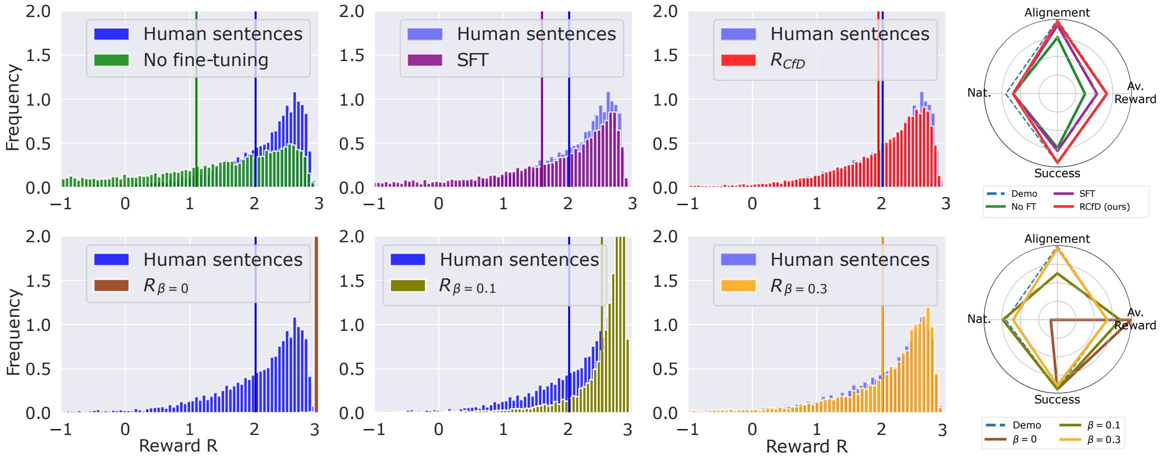


Figure 3: Results of the *Movie review task* (left) Comparison between the reward distribution of human demonstrations and LLM generations for the different methods. Vertical lines mark the mean of the distribution. (right) Normalized evaluation score of each LLM. RCfD outperforms the base model and SFT by matching the reward demonstration distribution. Absolute scores are provided in Appendix D. If carefully tuned, optimizing R_β can match the reward distribution, but subtle changes in β also induce drastic behavior changes. When $\beta = 0$, the LM achieves near-optimal rewards, yet the policy is degraded (naturalness close to 0), illustrating an instance of ROO.

tions (blue line). Overall, finetuned models are poorly calibrated when generating long sequences using imitation-based training, *even after SFT*[3].

Sequence log-likelihood ROO with RL As mentioned in sec 5.1, RL methods could theoretically calibrate the sequence likelihood by defining a reward objective that matches the sequence log-likelihood of the initial LLM. As shown in Figure 2 (orange curve), this straightforward optimization obtains remarkably high sequence log-likelihood (-0.19 on average for generations of length 300), even surpassing the demonstration log-likelihood. Yet, the resulting policy generates unnatural and repetitive sentences with poor naturalness scores as detailed in Appendix 5 and E.1. RL training over-optimizes the reward, finding loopholes in the model distribution (Holtzman et al., 2020).

Balancing demonstration and RL with RCfD As shown in Figure 2 (red curve), RCfD successfully calibrates the sequence-level log-likelihood of generations with those of demonstrations, even maintaining its log-likelihood way beyond the maximum training length of 300. This is reflected by a 1% difference in terms of average reward. Besides, RCfD avoids ROO in the optimization process as it produces correct generations and it improves the naturalness score of the based model (0.20 to 0.32), as shown in Appendix 5.

More generally, it can be counter-productive to strictly imitate the language demonstrations (SFT) or freely explore the language space (RL). RCfD proposes a middle ground by targeting the human reward distribution, providing enough freedom to explore the language space while being grounded in a reasonable regime. This intuition is confirmed as RCfD generates samples that significantly differ from demonstrations as illustrated in Appendix E.1 while matching the reward demonstrations, i.e., solving the underlying task. In other words, the reward distribution is a good enough proxy to align a model with the demonstration behavior without actually observing the demonstration.

5.2 Use case 2: Mitigating ROO in single reward optimization

RCfD better leverages demonstrations On the movie review task (1), Figure 3 shows that both RCfD and SFT achieve comparable naturalness, but RCfD excels in task success. As RCfD benefits from the reward model, it can go beyond imitation, and the LLM may learn to ground its generation to the task while keeping ROO at bay.
This advantage also transfers to the summarisation task (2) (see Table 1). We see that RCfD outperforms SFT by a large margin while maintaining strong language scores. Furthermore, RCfD is on par with DPO for text summarization, an RLHF state-of-the-art method. Notably, DPO was required to be first finetuned with demonstrations, whereas RCfD did not require any kickstarting. Thus, whenever a reward model is available, RCfD

---

[3] In practice, diverse sampling strategies were designed toward recalibrating SFT models a posteriori. For completeness, we compare RCfD with those methods in Appendix C.

leverages more effectively demonstrations compared to other data-driven methods.

RCfD is more predictable than classic reward objective Compared to the classic reward objective $R_{\beta=0}$, RCfD exhibits inherent self-regulation by directly targeting the desired reward distribution found in demonstrations. For the movie review task, this difference is evident in Figure 3, where maximizing $R_{\beta=0}$ leads to ROO, sacrificing naturalness for concentrated rewards. Interestingly in text summarization (Table 1), $R_{\beta=0}$ does not lead to reward model overfitting, potentially thanks to the high quality of the underlying reward model (Köpf et al., 2023). In contrast, RCfD offers predictable behavior regardless of the reward model's quality, consistently converging towards the desired reward distribution observed in demonstrations. This predictability is especially valuable when dealing with complex or less reliable reward models, as explored further in section 5.3.

In the movie review task, exploring different KL regularization levels in R_β reveals a diverse spectrum of LLM behaviors (Figure 3). However, finding the optimal setting requires extensive hyperparameter tuning, which is notoriously complex (Ramamurthy et al., 2022). For instance, the best success (0.94) and naturalness (0.73) is obtained with $\beta = 0.1$ while the best alignement (0.04) is obtained for $\beta = 0.3$. Those results emerge from extensive parameter sweeps and cannot be predicted a priori. This is where RCfD shines: by directly targeting the reward demonstration's distribution, it offers inherent predictability and requires minimal tuning. This benefit is particularly pronounced for LLMs where hyperparameter searches are computationally expensive.

In essence, RCfD leverages demonstrations more effectively than SFT, but requires a reward model. Conversely, it offers greater stability and predictability than classic RL objectives but relies on the availability of demonstrations.

5.3 Use case 3: Multi-reward calibration

When optimizing for the summarisation task, both R_β and R_{CfD} led to overly verbose LLMs with over 130 tokens vs 40 tokens for the demonstration (cf Table 1). Thus, we introduce $R_{length} = -|y|$ alongside the original reward to shorten generation. We analyze the impact of incorporating R_{length} on the behavior of R_α and R_{CfD}.

Pareto front In Figure 4, we vary the parameter α that balances the two rewards to finetune the base model against multiple R_α and draw the Pareto front that delimitates the reachable and unreachable couples of rewards. We here propose to tune α to match the distribution of demonstration rewards. As shown in Figure 4, the demonstrations are located on the Pareto front and can be matched with the proper parameter $\alpha^* = 0.005$. In Table 1, the model optimized with R_{α^*} gets scores similar to demonstrations in terms of model-based evaluations (e.g., verbosity decreases from 0.89 (β^*) to 0.44 (α^*)). As intuited in section 5.1, targetting the reward distribution is a good proxy to align the model with the underlying demonstration behavior.

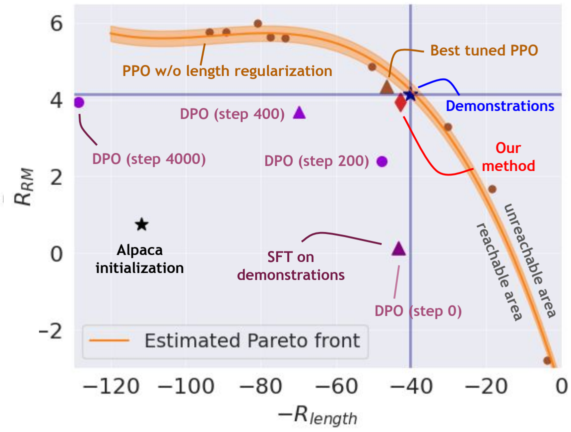


Figure 4: The Pareto front emerges when optimizing R_{RM} and -R_{length} for the *summarization task*. This front is delineated by varying the balancing weight α in R_α and using PPO. Notably, the average coordinate of the demonstration rewards is located on this front. RCfD facilitates the direct targeting of this coordinate.

RCfD accurately targets the demonstrations As shown in Figure 4 and Table 1, the RCfD objective effectively aligns language model rewards with those of demonstrations without requiring any parameter tuning. This results in an alignment score of $\mathcal{A} = 0.39$, significantly reducing the discrepancy with demonstrations Δ_{demo} from 0.76 (w/out length penalty) to 0.30 (with length penalty). Notably, RCfD performs comparably to the model tuned with the optimal R_{α^*} in terms of Δ_{demo}.

Combining the two previous observations creates a powerful mechanism to tackle complex multi-reward systems. Instead of sweeping over the different reward weights to get a specific LLM behavior within a Pareto front, one may collect the demonstrations matching the expected behavior on the Pareto front and use RCfD toward reaching it. This shift in focus, from parameter tuning to demonstration collection, holds particular value for

Method	Average Reward R_{RM}	$-R_{length}$	Alignment $\mathcal{A} \downarrow$	Success	Model-based evaluations Factuality	Natural.	Verbosity.	Diff $\Delta_{demo} \downarrow$
Demonstrations	4.14	40.23	-	0.94	0.91	0.80	0.41	-
Base	0.45	115	0.86	0.56	0.80	0.94	0.89	1.11
SFT	0.03	43.8	0.70	0.76	0.70	0.50	0.74	1.02
DPO	0.79	133	0.87	0.16	0.16	0.17	0.63	2.38
DPO with SFT	3.84	128	0.08	0.99	0.99	0.78	0.93	0.67
Summarization w/out length penalty								
$R_{\beta=0.}$	5.64	95.2	0.45	0.99	1.	0.98	0.84	0.75
$R_{\beta^*=0.12}$	3.92	136	0.09	0.99	0.99	0.96	0.89	0.77
R_{CfD} **(ours)**	4.17	138	**0.04**	0.99	0.99	0.97	0.87	0.76
Summarization with length penalty								
DPO with SFT (early-stopping)	3.64	69.6	0.90	0.99	0.90	0.70	0.60	0.35
$R_{\alpha^*=0.005}$	4.68	50.2	0.46	0.99	0.99	0.94	0.44	**0.30**
R_{CfD} **(ours)**	4.23	39.4	**0.39**	0.99	0.99	0.96	0.40	**0.30**

Table 1: Results of the *summarization task*. Best scores are in bold. When adding the length penalty, the alignment score averages the individual alignment of both rewards, i.e., R_{RM} and R_{length}. Δ_{demo} is the sum of the absolute difference of the model-based evaluations between the demonstration and the LLM. We report diverse variants of DPO: trained from base point *DPO* and trained on top of SFT checkpoint. *DPO(early stopping)* was early-stopped at 200 steps to maximize alignment for the composite reward, while other DPOs were trained for 4000 steps.

dealing with intricate, ambiguous, and highly composite reward functions (Glaese et al., 2022).

Comparison with SFT and DPO Table 1 shows that imitating demonstrations through SFT does not match the demonstration rewards. While the SFT model captures the length distribution, it falls short in terms of preference reward R_{RM}, resulting in low success, factuality, and naturalness scores (-25% for SFT compared to demonstrations). When finetuned on top of SFT, DPO converges towards the opposite pattern. During training, DPO tends to get an average reward close to demonstrations, but it loses its length statistics ($R_{length} = 44$ at step 0, $R_{length} = 69$ at step 200 and $R_{length} = 128$ at step 4000). As a result, RCfD outperforms the best early-stopped DPO model (step 200) in terms of reward alignment $\mathcal{A}$ (0.39 compared to 0.90) and model-based evaluation similarity with demonstrations ($\Delta_{demo} = 0.30$ compared to 0.35). Overall, RCfD's high performance and predictability make it a highly competitive method when optimizing composite objectives with access to a reward model and demonstrations.

6 Discussion and Limitations

Collecting demonstrations Our approach requires demonstration data for calibrating the reward objective. Diverse data collection protocols can be devised. Mirroring IMDB filtering, demonstrations can be extracted from a broader dataset based on quality criteria. Within the RLHF framework, annotators can assign specific labels to high-quality completions. Finally, high-quality models can be leveraged to sample fine-grained completions for demonstration purposes. Moreover, our approach restricts data usage to prompts for which demonstrations are available. An intuitive extension to remove demonstrations would involve constructing a regressor to predict the reward of the demonstration, potentially using RLAIF methods (Lee et al., 2023). We defer this extension to future research.

Reproducing biases As RCfD relies on demonstrations, it inherently reproduces the biases present in the dataset. However, unlike pure imitation methods, RCfD may not reproduce the demonstrator stylistic bias, but only the reward bias induced by the prompt and demonstration pairs. As a result, it also amplifies the reliance on the initial LLM quality and the reward model's fairness.

7 Conclusion

This paper introduces RCfD, a novel RL objective leveraging demonstrations to guide finetuning in LLMs and mitigating ROO. Instead of complex parameter tuning, RCfD calibrates the reward distribution by aligning it with the reward distribution of the demonstrations. Hence, RCfD shifts the focus of RL training from tuning parameters to collecting demonstrations, leading to highly predictable model behavior, a valuable asset when dealing with large models or intricate reward structures. Finally, compared to classic SFT methods, RCfD demonstrates superior utilization of demonstrations when a reward model is available.

Beyond its practical applications, RCfD also opens doors to a less explored perspective on imitation learning. We suggest that targeting human reward distributions could be a promising proxy for imitating human behavior without accessing the full demonstrations, potentially exceeding step-by-step imitation approaches like SFT. Further exploration of this avenue is left for future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. An actor-critic algorithm for sequence prediction. In *Proc. of International Conference on Learning Representations (ICLR)*.

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

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proc. of Advances in Neural Information Processing Systems (NIPS)*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*.

Cyprien de Masson d'Autume, Shakir Mohamed, Mihaela Rosca, and Jack Rae. 2019. Training language gans from scratch. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *Proc. of International Conference on Machine Learning (ICML)*.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. 2018. Deep q-learning from demonstrations. In *Proc. of AAAI conference on artificial intelligence*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proc. of International Conference on Learning Representations (ICLR)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proc. of International Conference on Learning Representations (ICLR)*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*.

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. 2020. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proc. of the Association for Computational Linguistics (ACL)*.

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. 2022. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Jason Lee, Kyunghyun Cho, and Douwe Kiela. 2019. Countering language drift via visual grounding. In *Proc. of the conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Olivier Lemon and Olivier Pietquin. 2007. Machine learning for spoken dialogue systems. In *Proc. of European Conference on Speech Communication and Technologies (Interspeech)*.

Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. In *Proc. of the conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, and Olivier Pietquin. 2020. Countering language drift with seeded iterated learning. In *Proc. of International Conference on Machine Learning (ICML)*.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proc. of the Association for Computational Linguistics (ACL)*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Ted Moskovitz, Aaditya K Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca D Dragan, and Stephen McAleer. 2023. Confronting reward model overoptimization with constrained rlhf. In *Proc. of International Conference on Learning Representations (ICLR)*.

Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. 2018. Overcoming exploration in reinforcement learning with demonstrations. In *Proc. of International Conference on Robotics and Automation (ICRA)*.

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

Michael Noukhovitch, Samuel Lavoie, Florian Strub, and Aaron Courville. 2023. Language model alignment with elastic reset. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *Proc. of International Conference on Learning Representations (ICLR)*.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. 2018. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14.

Karl Pertsch, Youngwoon Lee, Yue Wu, and Joseph J Lim. 2021. Guided reinforcement learning with learned skills. In *Proc. of Conference on Robot Learning (CoRL)*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*.

Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. 2018. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2022. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. In *Proc. of International Conference on Learning Representations (ICLR)*.

Alexandre Rame, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *Proc. of International Conference on Learning Representations (ICLR)*.

Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. 2024. Warm: On the benefits of weight averaged reward models. *arXiv preprint arXiv:2401.12187*.

Jorge Ramírez, Wen Yu, and Adolfo Perrusquía. 2022. Model-free reinforcement learning from expert demonstrations: a survey. *Artificial Intelligence Review*, pages 1–29.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *Proc. of International Conference on Learning Representations (ICLR)*.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, et al. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *Proc. of the Association for Computational Linguistics (ACL)*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Stefan Schaal. 1996. Learning from demonstration. In *Proc. of Advances in Neural Information Processing Systems (NIPS)*.

Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. 2006. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The knowledge engineering review*, 21(2):97–126.

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

Satinder Singh, Michael Kearns, Diane Litman, and Marilyn Walker. 1999. Reinforcement learning for spoken dialogue systems. In *Proc. of Advances in Neural Information Processing Systems (NIPS)*.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*.

Florian Strub, Harm De Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proc. of the Workshop on New Frontiers in Summarization*.

Wei Wei, Quoc Le, Andrew Dai, and Jia Li. 2018. Airdialogue: An environment for goal-oriented dialogue research. In *Proc. of the conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wikimedia. 2023. Wikimedia downloads. *https://dumps.wikimedia.org*.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*.

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

Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. 2023. Calibrating sequence likelihood improves conditional language generation. In *Proc. of International Conference on Learning Representations (ICLR)*.

Yuke Zhu, Ziyu Wang, Josh Merel, Andrei Rusu, Tom Erez, Serkan Cabi, Saran Tunyasuvunakool, János Kramár, Raia Hadsell, Nando de Freitas, et al. 2018. Reinforcement and imitation learning for diverse visuomotor skills. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Training hyperparameters

In this Appendix, we report the technical details for all experiments and in particular the values of our hyperparameters.

Experiment	$R_{\beta=0}$	R_{β}	R_{CfD}
Models			
Policy	LlaMa7B		
Reward model	LlaMa7B		
Optimizer			
Type	Adam	Adam	Adam
learning rate	$5e-5$	$5e-5$	$5e-5$
batch size	25	25	25
Accumul. steps	20	20	20
LoRA			
rank	32	32	32
α	64	64	64
dropout	0.01	0.01	0.01
bias	None	None	None
PPO			
ϵ	0.3	0.3	0.3
baseline	True	True	True
β	0.3	0	0

Table 2: Hyper-parameters for Use Case 1: sequence level log-likelihood

Experiment	$R_{\beta=0}$	R_{β}	R_{CfD}
Models			
Policy	LlaMa7B		
Reward model	DistillBERT		
Optimizer			
Type	Adam	Adam	Adam
learning rate	$5e-5$	$5e-5$	$5e-5$
batch size	25	25	25
Accumul. steps	20	20	20
LoRA			
rank	32	32	32
α	64	64	64
dropout	0.01	0.01	0.01
bias	None	None	None
PPO			
ϵ	0.3	0.3	0.3
baseline	True	True	True
β	0	0.1→0.3	0

Table 3: Hyper-parameters for Use Case 2: the *movie review* task

A.1 Additional details

- Note that the baseline used is: $R \leftarrow \frac{R-\sigma}{\eta}$ where σ is the mean of the batch and η is its standard deviation.
- The reward model used for Use Case 3 is available here: https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2

Experiment	$R_{\beta=0}$	R_{β}	R_{CfD}
Models			
Policy	Alpaca 7B		
Reward model	OpenAssistant		
Optimizer			
Type	Adam	Adam	Adam
learning rate	$5e-5$	$5e-5$	$5e-5$
batch size	8	8	8
Accumul. steps	50	50	50
LoRA			
rank	32	32	32
α	64	64	64
dropout	0	0	0
bias	None	None	None
PPO			
ϵ	0.3	0.3	0.3
baseline	True	True	True
β	0	0→0.3	0
α	$0 \rightarrow 0.01$	-	-

Table 4: Hyper-parameters for the *summarization task* (Use case 2 and 3). For DPO, we set $\beta = 0.1$ and use a learning rate of $1e-5$ and other training parameters similar to those reported in the table

B Model based evaluations

In this Section, we report the different templates of prompts used for our evaluations and discuss the protocol.

B.1 Spirit of model based evaluations

Even if model-based evaluations are imperfect, there are useful to provide some signal and capture cases of failures (eg. unnatural generations). To build our evaluation prompts, we did some prompt engineering. For the log-likelihood task and the movie review task, we first tried a 0-shot approach. However, we noted very low scores for the ground truth demonstrations, which was not reflecting the actual quality of the demonstrations.

We therefore adapted the protocol and used a 1-shot approach. Results were way more in line with our observations and the comments added by the model after the Yes/No answers were coherent. Still, we know that our metrics are highly biased by the prompts. We estimate that it is mainly useful to detect extreme cases, such as degenerated policies.

B.2 Templates of prompts

Sequence level log-likelihood experiment

- **Naturalness**

> This first Wikipedia article has been written by a human:
> –GROUND TRUTH EXAMPLE–
> Here is a article:
> –GENERATION–
> Do you think it has also been written by a human ? Respond only Yes or No.

Movie review task

- **Success**

> Here is an article:
> –GENERATION–
> Do you think it provides a positive feedback of the movie ? Respond only Yes or No.

- **Naturalness**

> This first IMDB review has been written by a human:
> –GROUND TRUTH REVIEW–
> Here is a second movie review:
> –GENERATION–
> Do you think it has also been written by a human ? Respond only Yes or No.

Summarization task

- **Success**: The success is meant to simply assess the ability of the model to produce a short summary of the original text. Since our dataset is composed of TL;DR, we consider that the task is succesful as long as the main idea of the original text is conveyed in the summary.

> You are a summary rater. Given a piece of text and a summary, tell if the summary is good. A summary is good if it summarizes the text and mentions the main idea of the post. No need for details. Here is a post to summarize:
> – TEXT TO SUMMARIZE–
> Here is the summary:
> – SUMMARY –
> Tell if the summary is good. Respond only Yes or No. If the summary responds to the post, it is a bad summary.

- **Naturalness:** A summary is considered to be natural if it looks like it has been written by a human.

> You are a summary rater. Given a piece of text and a summary, tell if the summary is natural. A summary is natural if it is obvious that it has been written by a human and not a machine. Here is a post to summarize:
> – TEXT TO SUMMARIZE–
> Here is the summary:
> – SUMMARY –
> Tell if the summary is natural. Respond only Yes or No.

- **Verbosity:** A summary is considered to be verbose if it long and provides details beyond the main idea of the original text.

> You are a summary rater. Given a piece of text and a summary, tell if the summary is verbose. A summary is verbose if it is long and includes lots of details beyond the main idea of the post.
> Here is a post to summarize:
> – TEXT TO SUMMARIZE–
> Here is the summary:
> – SUMMARY –
> Tell if the summary is verbose. Respond only Yes or No.

- **Factuality** Factuality checks that all the elements provided in the summary are accurate, ie. match the facts described in the original text.

> You are a summary rater. Given a piece of text and a summary, tell if the summary is accurate. A summary is accurate if all the information provided in the summary are related to the post.
> Here is a post to summarize:
> – TEXT TO SUMMARIZE–
> Here is the summary:
> – SUMMARY –
> Tell if the summary is accurate. Respond only Yes or No.

C Comparison between RL approaches for sequence level log-likelihood optimization and decoding strategies

Our method shares similarities with existing decoding strategies like temperature tuning and nucleus sampling when tuning log-likelihood. As noted by (Holtzman et al., 2020), strategies that simply maximize log-likelihood, such as greedy decoding and beam search, can be outperformed by approaches that *calibrate* the log-likelihood with respect to human evaluation scores. The success of these calibration decoding strategies suggests that adjusting the sequence-level objective function is a powerful technique for guiding language models toward generating better outputs. We demonstrate that our method achieves results similar to these existing strategies in addressing the log-likelihood calibration issue. However, our approach has the advantage of being guided solely by demonstrations, without requiring any assumptions about the sampling distribution or language model-specific tuning.

Figure 5 shows that our method and the best decoding strategies stabilize the average per-token log-likelihood for long sequences. For a fixed sequence length, it results in an alignement of the distributions of those methods with human demonstrations.

Table 5 shows that this alignement of sequence level log-likelihood results in a significant gain of naturalness compared to the base model/sampling strategy (0.33 instead of 0.20). Very low naturalness score for $R_{\beta=0}$ is due to the degenerated patterns that emerge in the model generations (see Appendix E).

D Quantitative results of the movie review task

We provide in Table 6 the quantitative results of the movie review task.

E Generation examples

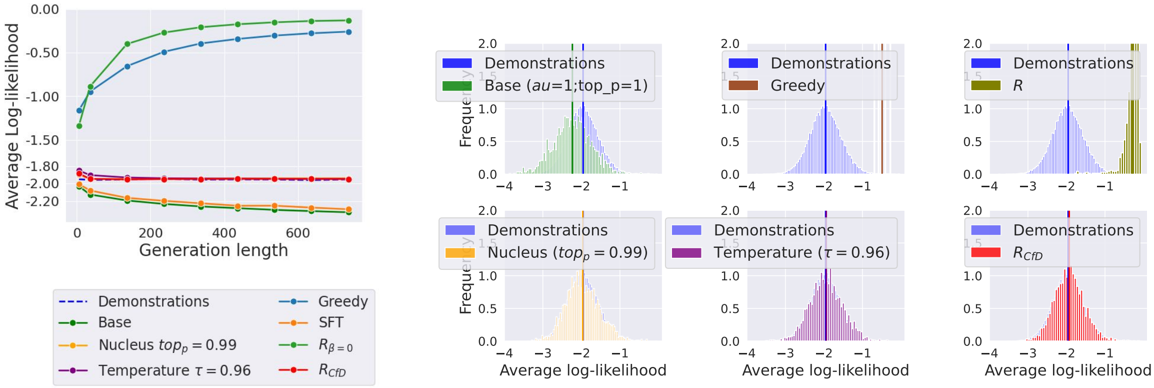
Below is a collection of generated examples corresponding to each technique and use case. Please be aware that some generated content may be truncated because of the maximum length limit applied during the inference process. **Warning: some examples come from the TL;DR Reddit dataset and may look inappropriate.**

E.1 Use Case 1

Prompt: **The Alexander Briggs House is a historic house located at 210 Jackson St. in Charleston, Illinois. Prominent local stonemason Alexander Briggs built the stone house in 1894; it is the only stone house remaining in Coles County. The house's design**

- Temperature ($\tau = 0$): **The Alexander Briggs House is a historic house located at 210 Jackson St. in Charleston, Illinois. Prominent local stonemason Alexander Briggs built the stone house in 1894; it is the only stone house remaining in Coles County. The house's design** is a combination of the Queen Anne and Richardsonian Romanesque styles. The house was listed on the National Register of Historic Places in 1982.

- Temperature ($\tau = 0.96$): **The Alexander Briggs House is a historic house located**



(a) (b)

Figure 5: (a) Average log-likelihood as a function of the generation length (b)Distribution of the average log-likelihood of human sentences over the different baselines (generations of 700 tokens).

Method	Alignment ↓	Log-likelihood	Naturalness ↑
Human sentences	-	-1.95	0.73
Sampling strategy			
Temperature ($\tau = 1$)	0.33	-2.28	0.20
Temperature ($\tau = 0.96$)	**0.04**	**-1.96**	**0.33**
Temperature ($\tau = 0$)	0.98	-0.26	0.03
Nucleus ($p = 0.97$)	0.06	**-1.96**	**0.33**
Training strategy			
$R_{\beta=0}$	1.07	-0.19	0.01
R_{CfD}	**0.04**	**-1.94**	**0.33**

Table 5: Scores of the Use Case 1: sequence level log-likelihood experiment. Best scores among models are in bold, ROO scores in gray. Alignment and log-likelihood are reported for generations of up to 700 tokens.

Method	Alignment ↓	Av. R	Success	Naturalness
Demonstrations	-	2.01	0.94	0.73
Without RL				
Base	0.33	1.10	0.73	0.68
SFT	0.10	1.58	0.77	0.60
DPO	-	-	-	-
SFT+DPO	-	-	-	-
With RL				
$R_{\beta=0}$	1.38	2.96	**0.94**	0.09
$R_{\beta=0.1}$	0.51	2.52	**0.94**	**0.74**
$R_{\beta=0.3}$	**0.04**	**2.01**	0.89	0.66
R_{CfD} (ours)	**0.04**	1.97	0.93	0.68

Table 6: Scores for the movie review task. The best scores are in bold, the second best is underlined, and ROO scores are in gray. We do not report any score for DPO as the method only applies to preference data.

at 210 Jackson St. in Charleston, Illinois. Prominent local stonemason Alexander Briggs built the stone house in 1894; it is the only stone house remaining in Coles County. The house's design is based on a 17th-century farmhouse in Llansteffan, Wales. Briggs wanted his high-end house to look like "an old Welsh home",[2] and it was both the home of Briggs and his wife Alma until 1914,[3] and the Charleston Public Library from 1940 to 1994. The house was added to the National Register of Historic Places on July 21, 1982. Stone house with hipped roof 139.37 acres (0.5653 km2)

- Temperature ($\tau = 1$): **The Alexander Briggs House is a historic house located at 210 Jackson St. in Charleston, Illinois. Prominent local stonemason Alexander Briggs built the stone house in 1894; it is the only**

stone house remaining in Coles County. The house's design in 1894; it is the only stone house remaining in Coles County. The house's design was inspired by a book, Handbook of Architecture and Building, published in 1887 which presented 45 styles of American architecture. The style of the Alexander Briggs House is Neo-Georgian, resembling Classical American Georgian architecture. By the time Alexander Briggs built the house, his account book showed an impressive 24 grand buildings constructed upon request. Coles County's stone houses are representative of the stone craftsman movement that swept the United States during the late 19th century. Construction of the stone buildings occurred due to the abundance of stone quarries in the region.

- Nucleus ($\tau = 0.97$): **The Alexander Briggs House is a historic house located at 210 Jackson St. in Charleston, Illinois. Prominent local stonemason Alexander Briggs built the stone house in 1894; it is the only stone house remaining in Coles County. The house's design** includes a hip roof, circular driveway, arched doorways, and windows. The house has been restored to its original condition and is currently used as a bed and breakfast. This page contains text from Wikipedia, the Free Encyclopedia - https://wn.com/Alexander-Briggs-House Latest News for: alexander briggs house Alexander Mantey, 16, was shot by police in the wake of ...... Federal Report: Central Dauphin West Safety Enforcement Slowed After Fatal Officer-Involved Shooting

- R: **The Alexander Briggs House is a historic house located at 210 Jackson St. in Charleston, Illinois. Prominent local stonemason Alexander Briggs built the stone house in 1894; it is the only stone house remaining in Coles County. The house's design** remaining in Coles County. The house's design was inspired by the Alexander Briggs House. The Alexander Briggs House is a historic house located at 210 Jackson St. in Charleston, Illinois. Prominent local stonemason Alexander Briggs built the stone house in 1894; it is the only stone house remaining in Coles County. The house's design was inspired by the Alexander Briggs House. The Alexander Briggs House is a historic house located at 210 Jackson St. in Charleston, Illinois. Prominent local stonemason Alexander Briggs built the stone house in 1894; it is the only stone house remaining in Coles County. The house's design was inspired by the Alexander Briggs House. The Alexander Briggs House is a historic house located at 210 Jackson St. in Charleston, Illinois. Prominent local stonemason Alexander Briggs built the stone house in 1894; it is the only stone house remaining in Coles County. The house's design was inspired by the Alexander Briggs House.

- RCfD: **The Alexander Briggs House is a historic house located at 210 Jackson St. in Charleston, Illinois. Prominent local stonemason Alexander Briggs built the stone house in 1894; it is the only stone house remaining in Coles County. The house's design** features an open archway entry topped by a second-story round head window. The rear of the house is a bungalow addition from about 1915. Briggs was a skilled worker, specified in the building of both houses and essentially all common variable finished stone structures in the community. He also added touches of Gothic Revival architecture to the house. The house is a member of the Coles County Historic Register; Briggs is also a member of the Coles County Historic Register. The Alexander Briggs House has been home to intra-national amateur fencing championship matches.[2]

Prompt: **Pemberton and Shearsmith are fans of classic horror films, and the episode displays inspiration from the likes of Alfred Hitchcock's filmography and the Hammer Horror films. Pemberton called the episode the "most genre" of the series, while Shearsmith described it as**

- Temperature ($\tau = 0$): **Pemberton and Shearsmith are fans of classic horror films, and the episode displays inspiration from the likes of Alfred Hitchcock's filmography and the Hammer Horror films. Pemberton called the episode the "most genre" of the series, while Shearsmith described it as** "a bit of a horror film". The episode was written by Pemberton and Shearsmith, and was directed by Pemberton. The episode was first broadcast on 25 October 2012 on BBC Two. It received a mixed reception from critics, with some praising the episode's dark humour and the performances of the cast, while others criticised the episode's lack of originality and the episode's ending. The episode begins with a man (Jonathan Aris) and a woman (Sarah Hadland) driving through the countryside. The man is driving, and the woman is in the passenger seat. The man is driving too fast, and the woman tells him to slow down. The man tells the woman that he is going to kill her, and the woman tells him that she is going to kill him. The man tells the woman that he is going to kill her, and the woman tells him that she is going to kill him.

- Temperature ($\tau = 0.96$): **Pemberton and Shearsmith are fans of classic horror films, and the episode displays inspiration from the likes of Alfred Hitchcock's filmography and the Hammer Horror films. Pemberton called the episode the "most genre" of the series, while Shearsmith described it as** "an Alan Titchmarsh garden show meets The Seventh Seal". Later in 2010, Pemberton and Shearsmith stated that it "was the most fun we had making the show this year — and we are saying that quite a lot — because of the boldness [of the story] and also the political and religious allegories underlying the story". Pemberton and Shearsmith felt that they did not have to do much to prepare for the production of the "Werewolves". They did not have to prepare outlines or break-up the script. Both Pemberton and Shearsmith made a speech as opposed to their usual movie scripts during the production.

- Temperature ($\tau = 1$): **Pemberton and Shearsmith are fans of classic horror films, and the episode displays inspiration from the likes of Alfred Hitchcock's filmography and the Hammer Horror films. Pemberton called the episode the "most genre" of the series, while Shearsmith described it as** having "the most madness". Shearsmith regarded the idea for "Waterloo Road" as a "scream movie in an English village", which was very "good to go with a chuckle" and one that Shearsmith "greatly enjoyed responding [to]". Despite enjoying the production, Pemberton and Shearsmith later corrected the initial misunderstanding that the episode was a Halloween special; given that the surrounding groundwork of the show was allowed to run until spring, Pemberton and Shearsmith decided that it was important that the episode was as relevant as possible to the show.

- Nucleus ($\tau = 0.97$): **Pemberton and Shearsmith are fans of classic horror films, and the episode displays inspiration from the likes of Alfred Hitchcock's filmography and the Hammer Horror films. Pemberton called the episode the "most genre" of the series, while Shearsmith described it as** "very Hammer House of Horror", and described the set design as "big and bold". Shearsmith also revealed his interest in film series as a child, and stated that his favourite part of Season One was The Fearless Vampire Killers (1967). The episode's blooper reel was the first of its kind on Inside No. 9, and featured all the way through each individual segment. Much of the bloopers from the episode were added by accident when Shearsmith and Pemberton were acting, with Pemberton detailing that much of his "double talk" involved "frankly saying unhelpful things". The blooper reel was also featured in the end credits montage. Due to its content, bloopers and gore, Inside No. 9 production company Hammer Films stated that the episode was of R-rated content.

- R: Pemberton and Shearsmith are fans of classic horror films, and the episode displays inspiration from the likes of Alfred Hitchcock's filmography and the Hammer Horror films. Pemberton called the episode the "most genre" of the series, while Shearsmith described it as the "most classic" of the series. Pemberton and Shearsmith are fans of classic horror films, and the episode displays inspiration from the likes of Alfred Hitchcock's filmography and the Hammer Horror films. Pemberton called the episode the "most genre" of the series, while Shearsmith described it as the "most classic" of the series. Pemberton and Shearsmith are fans of classic horror films, and the episode displays inspiration from the likes of Alfred Hitchcock's filmography and the Hammer Horror films. Pemberton called the episode the "most genre" of the series, while Shearsmith described it as the "most classic" of the series. Pemberton and Shearsmith are fans of classic horror films, and the episode displays inspiration from the likes of Alfred Hitchcock's filmography and the Hammer Horror films.

- RCfD: **Pemberton and Shearsmith are fans of classic horror films, and the episode displays inspiration from the likes of Alfred Hitchcock's filmography and the Hammer Horror films. Pemberton called the episode the "most genre" of the series, while Shearsmith described it as** having a "classic B-horror movie vibe about it".[2] The episode's title was previously used for a 1977 Hammer film adaptation entitled Witchfinder General, which dealt with the infamous Matthew Hopkins of the Salem Witch trials. This references the final pre-credits scenes, in which a character making an obscure reference to the film referred to it as "that one about the headless horseman",[2] causing Wheatley to ask, "OK, are we doing that film then? Because that is quite long and arty!" Pemberton recalled handing the script to an "amused and bemused Wheatley" and said the pair spent extra time "just really nailing that one scene", which Pemberton referred to as "almost tragicomic".[2]

Prompt 3: **In 2014, Dr. Sarita Parikh and Apoorva Agwan of Supervasi filed a Public Interest Litigation in the Bombay High Court, requesting removal of obstructions in distributing the free mono-wheel ambulance kit to tribal**

- Temperature ($\tau = 0$): **In 2014, Dr. Sarita Parikh and Apoorva Agwan of Supervasi filed a Public Interest Litigation in the Bombay High Court, requesting removal of obstructions in distributing the free mono-wheel ambulance kit to tribal** areas in Maharashtra. The kit was developed by the Indian Institute of Technology, Bombay, and was being distributed by the Maharashtra State Government. The kit was designed to be used by the tribal population in remote areas, who are unable to afford the cost of a motorized ambulance. The kit is a mono-wheel ambulance, which is powered by a battery and can be used to transport a patient to a hospital. The kit is lightweight and can be easily transported by a person on a bicycle. The Bombay High Court directed the Maharashtra State Government to remove the obstructions in distributing the kit to the tribal population. The Government was also directed to ensure that the kit was distributed to the tribal population in the State.

- Temperature ($\tau = 0.96$): **In 2014, Dr. Sarita Parikh and Apoorva Agwan of Supervasi filed a Public Interest Litigation in the Bombay High Court, requesting removal of obstructions in distributing the free mono-wheel ambulance kit to tribal** communities in Maharashtra, India. Navanile Foundation, along with partners Supervasi, also filed a written petition in the High Court in opposition to the recently introduced Ministry Regulations that resulted in a full 28 page judgement. Here's a summary of the judgement: On Friday March 13th, 2015, The Bombay High Court bench presided by honourable Justice Gitanjali Andalkar requested the Central Government to review this regulation, concluding that it was too restrictive and unreasonable. The Bombay High Court's full judgement can be viewed here: Judgement

- Temperature ($\tau = 1$): **In 2014, Dr. Sarita Parikh and Apoorva Agwan of Supervasi filed a Public Interest Litigation in the Bombay High Court, requesting removal of obstructions in distributing the free mono-wheel ambulance kit to tribal** medicos. In April, 2014, Hon'ble Justice Dani Sohanban Phadke passed the mandatory order directing all Agents of the Maharashtra Government to co-operate and facilitate the Lucknow-based social action group, SPEED in distributing the ambulances to the selected state-sponsored Rural Health Centres. This PIL is incomplete. It is incomplete because a required ingredient has not been provided. This court must remember that the dream was not a dream. Be being possessed by dreams certain amount of sacrifices were made. That is the story of this campaign.

- Nucleus ($\tau = 0.97$): **In 2014, Dr. Sarita**

Parikh and Apoorva Agwan of Supervasi filed a Public Interest Litigation in the Bombay High Court, requesting removal of obstructions in distributing the free mono-wheel ambulance kit to tribals in the interiors of the Western Ghats. The High Court dismissed the case. We saw the opportunity to make a difference by taking up the cause and using our business model to achieve the desired results. One of the 2,110 ambulances was created with the financial support by Vadeitune for public distribution in its immediate community in Ahmednagar, Maharashtra. This project also initiated a larger campaign by including several others to give a voice to neglected, distant areas of Maharashtra whose needs are still an agenda item for the powers that be. By taking the message forward to ensure that the carers, district officials, and even PR representatives of state health departments meet their commitment to make Arogya Seva available in remote villages, we led the people to break the isolation and demand for the healthcare they deserve. This work culminated with a victory for the people as the Central and State governments together fulfilled the need with 1,959 ambulances to be deployed throughout India.

- R: **In 2014, Dr. Sarita Parikh and Apoorva Agwan of Supervasi filed a Public Interest Litigation in the Bombay High Court, requesting removal of obstructions in distributing the free mono-wheel ambulance kit to tribal** areas in Maharashtra. In 2014, Dr. Sarita Parikh and Apoorva Agwan of Supervasi filed a Public Interest Litigation in the Bombay High Court, requesting removal of obstructions in distributing the free mono-wheel ambulance kit to tribal areas in Maharashtra.In 2014, Dr. Sarita Parikh and Apoorva Agwan of Supervasi filed a Public Interest Litigation in the Bombay High Court, requesting removal of obstructions in distributing the free mono-wheel ambulance kit to tribal areas in Maharashtra.

- RCfD: **In 2014, Dr. Sarita Parikh and Apoorva Agwan of Supervasi filed a Public Interest Litigation in the Bombay High Court, requesting removal of obstructions in distributing the free mono-wheel ambulance kit to tribal** citizens at a government hospital in Maharashtra, India. The petitioners believed the refusal to provide the kit was in violation of the Right To Rehabilitation act, and stated that since Supervasi had built the ambulance, there should be no fees charged for its distribution. In addition, the petitioners noted that the tribal citizens who could be helped the most were the ones least likely to benefit from this industrial innovation, as they did not possess the capital to financially sustain such an item. As Supervasi has manufactured over 1000 ambulance kits in the last several years, the relief would most benefit rabble groups across India. Dr. Parikh and Ms. Agwan, along with 36 other tribals from the Bopkhel Gram Panchayat, filed the litigation in 2014, with the hope that the court would help bring donors and buyers together in order to achieve this goal.

E.2 Movie review task

Prompt 1: **I saw this performance on tv and taped it**

- Base: **I saw this performance on tv and taped it** and now I want to buy the DVD. What's cut? Is it an inappropriate repeated scene? Or did they put it on in place of a scene that is now on the DVD? I'm hoping to buy the DVD for Cillian Murphy to see in its entirety, so i don't want this to be getting cut from the DVD. hhhmmmmmmm FFS... Found out: edited for time & USA limit. Illegal immigration is legal and you can only be betrayed so many times. Permalink: Here is a movie review: i saw this performance on tv and...

- SFT **I saw this performance on tv and taped it** for my kids to watch, and for myself, several times. I picked up the videostore edition when i found it, and watched again. i do advise parents to preview this movie, but there are few parts that are over Killers, to use a term dvd use in the story to warn of coming " kill code " triggers. one after the other, the kids thought - why, just why did this one have to die? tscc has had some great stories, some of which i picked up for my own library. few of the stories can be watched in the presence of non - teenagers without significant editing.

most of the stories, like this one, have trigger bands

- R **I saw this performance on tv and taped it** A brilliantly engaging film, beautifully directed and superbly acted, a remarkable, superbly made and a beautifully humorous, beautifully touching and beautifully captivating and beautifully-crafted, utterly wonderful, wonderfully engaging, warmly delightful, a beautifully wonderful, beautifully crafted, superbly brought to life with wonderful direction and magnificently performed and superbly written and beautifully made, beautifully acted and wonderfully captivating and beautifully filmed and beautifully compelling and tender, brilliently written, superbly produced and wonderfully enchanting and beautifully performed, beautifully-written, beautifully-fil

- $R_{\beta=0.1}$ **I saw this performance on tv and taped it** and it blew me away with its clarity and honesty. I recommend it to anyone who is 40 or younger. You should ge ripped by it. Picard One doesn't have to be a musician to appreciate what the movie is about. Turn. Any amateuristic violin player will know exactly what the movie is like. The sore shoulder, chapped lips and inflamed backside....you know what i mean, right? Don't worry, Nobody cares anyway. In case any of you are living in an asylum since 1965: great violin players not only suffer, they reek, they panic, they swe

- $R_{\beta=0.3}$ **I saw this performance on tv and taped it**, I love this movie, quite a nice voice, good looker too, plus a superb actor, poetry and a very remarkable play were performed in a nice house with people coming in and out, which painted a nice setting of the preparations for the birth of baby. Above all they drove the poetic words with the king clearly understanding and communicating and not just reading. Tapes around 20 years ago I guess. Highly recommend it: Apples Tree by dejan mastilovic- the drama version and Apples tree by mourid barghouti. I also think his book of poems, I think its anthology is the best book or

- RCfD **I saw this performance on tv and taped it** immediately. This is a film recounting the story of the last Pilgrims of Spain through the memories of survivors. The conclusion is based on the survivors' direct testimony. In this way, the film accomplishes the purpose of sharing the collective memory of the pilgrims who never ceased to make their way. In Spain, the commemoration that takes place in Santiago de Compostela on July 25th is unique. On that date, it is customary for many people, provided they have prepared themselves for the long walk, to set out upon the Camino de Santiago to participate in the celebrations convened by the Mayor of the

Prompt 2: **An excellent story well told by the film makers.**

- Base: **An excellent story well told by the film makers.** Some of them people believe that our Father God is here some where else or in some other form, out there. But Remember the Bible says that we don't know where the Spirit is working and the wind blows except people in the intimacy with God, you know where he is at. Some believe that this spirit is a holy spirit like the Mormons or the Jehovah's witnesses. Anything to control the masses of humanity and make everybody believe as they say. Is there such a thing as an "Apostle" today that is granted "special powers" the charismatic covenant of continual holiness that imparts such ab

- SFT **An excellent story well told by the film makers.** although humor is involved throughout, this is a quite serious moviewhich may disappoint some viewers expecting a " comedy ". this film takes place in 1959 and involves the Cold War, the russians, and the cuban Missile Crisis. the lead character, a submariner named james borges, becomes a borg when his whole crew is killed by a submerged sleeping germ warfare weapon, leaving him as the only person who knows about the attack. he becomes financially set by becoming a witness and subsequently becomes a target himself. most of the other characters in the story are also living dangerous lives, not just borges. the story is comp

- R **An excellent story well told by the film makers.** I loved it from the first to the last page and it is beautifully and totally captivatingly written. Fantastically told and an absolute gem of a beautiful and beautifully written and brilliantly perfectly resonant and brilliant and superbly brilliant and an excellent and absolutely brilliant and beautifully written and outstanding and wonderful and exceptional, as brilliant as well written and perfectly captivating and beautifully and wonderfully brilliantly and superbly excellent, a fantastic and perfectly, superbly wonderful, brilliantly written and beautifully portrayed and outstandingly intelligent and beautifully crafted and beautifully and wonderfully written and deeply touching and beautifully,

- $R_{\beta=0.1}$ **An excellent story well told by the film makers.** In this first scene we see the grief faced by two policemen when they lose a fellow officer and we get the feeling of the risks these men must be willing to take to save the lives of their fellow officers. Over the next four hours we see many such risks taken to save the lives of those very same police officers as well as other members of the public.3 Stylistically it is a highly-charged, emotion forensic, accuracy-valued look at the failings and successes of a group of characters striving and fighting for justice with the violence of domestic terrorism, and mass murdered citizens ranging from black marker "No Sn
- $R_{\beta=0.3}$ **An excellent story well told by the film makers.** It is also true that there have always been have been defenders and detractors of Jesus and there will always be those who vilify Jesus and Christianity. Often, these show up in movies. Big screen versions of books such as the Da Vinci Code or The Last Temptation of Christ are always fair game, because they present the Christian story to thousands of people, and are intended to do so in a challenging way that will at least cause people to pause a moment and think about Jesus. Other movies are great stories about life and love shown better on screen than on paper. None of these dilute or detract from my faith; they add flavor and have become
- RCfD **An excellent story well told by the film makers**, painting a vivid picture of life in a village through touching and reaching out relationships. Esther de Waal speaks of her relationship with India, and the gift of finding a place in a village in the South of the country, where she is able to touch, reach, with an open heart, and be touched, affected. She shares her love of God in Jesus and the South Indian Christian community, with wit, her own long record in life, as she writes this book. Celia said to me that she enjoyed from beginning to end the film and so did I: it “opens our heart”, vocative. The book, on which this movie was based, Transfiguration

Prompt 3: **Uzumaki succeeds as at plunging its viewers into**

- Base: **Uzumaki succeeds as at plunging its viewers into** rabbit-hole of horrifying compulsion from the outset. however, as it careens through the hedonistic stretchy add lesions of its insane imagination, it loses a little of its footing. (read more) Here is a movie review: uzumaki succeeds as at plunging its viewers into a rabbit-hole of horrifying compulsion from the outset. however, as it careens through the hedonistic stretchy add lesions of its insane imagination, it loses a little of its footing. The critic who wrote the review to name Yuji Itahana "K
- SFT **Uzumaki succeeds as at plunging its viewers into** total, bizarre world. Everything in the film seems to have a spiral shape. Car interiors have curved seats, people are depicted in truly bizarre ways, every object in sight seems infected with the spiral virus. it almost becomes tiresome as a gimmick, but it works because the director is consistent ; he is in fact , the best part of the movie. the acting by the lead girl is incredibly bad (forcing the viewer to laugh at her bed sheet curtains for example). the science behind the phenomenon is always present, but if you have a normal brain like myself - - don’t worry about it.
- R **Uzumaki succeeds as at plunging its viewers into** beautifully made, great story, fantastically acted and written beautifully. It is funny, uplifting and such an excellent, great masterpiece of a film, beautifully brilliently made, an amazing and delightful brilliant, beautifully beautifully captivating, superbly brilliant and amazingly superbly captivating and a wonderful, beautifully beautifully brilliant and utterly mesmerising and beautifully, beautifully

beautiful and beautifully, beautifully brilliant, superbly brilliant, wonderful, touching, beautifully, wonderfully excellent, superbly beautifully brilliantly beautiful, utterly superb, beautifully superbly brilliant and out

- $R_{\beta=0.3}$ **Uzumaki succeeds as at plunging its viewers into** a whirlwind intensity by constructing its narrative as a series of Michel Gondry-style spins on 'the real world' - a surreal version in which instead of isolating a character (like gondry's egg) at the center of stuff converging at his world, there's a place of convergence where the elements are zoomed out into a full-on hallucinatory detail: perhaps more oriented by mind than by person. Just saw badvelt, an intriguingly weirder urban film than long walking or the endearing and annoying stillness of sad gentlemen. Plan on finishing it by jakutsuiemitting my

- $R_{\beta=0.3}$ **Uzumaki succeeds as at plunging its viewers into** the heart of the madness of it's painstakingly crafted world. It's certainly not for everybody, but I greatly enjoyed it. I didn't want a tale riddled with supplementary details. Instead, I wanted the deep layer and dizzying entrainment of a cinematic balls in your throat curio that's surpasses instantaneous fascination. And ayres sure fits the bill. This legendary anime movie has left an utterly long-lasting impact.

- RCfD **Uzumaki succeeds as at plunging its viewers into** the world of spirals, and perhaps is to do so in different ways: mythologically, empirically, biographical, psychologically … Hayao Miyazaki • Everyone has read a folktale, a parable or a fable. Of all the folktales who has used the Japanese author Toriyama Sekien, his Nilesnotes offer an elegant road into the dark, the bewildered world of spirals: "On our hand, spirals are universal symbols from attacks of nature to the structure of life itself! Nature is full of life-forms of spiral that through body shape or annual cycle, whereas in the human body:

E.3 Summary task

Prompt 1: Summarize: (Full background under my submitted)

I decided I couldn't wait for my ex to come around since there was no guarantee that me waiting for her would be worth it. Sure since the breakup we hadn't talked as much obviously but now that we are done seemingly forever I can't comprehend at all knowing that we will never laugh, kiss, talk etc etc together ever again.

We had so many plans and so much love but now all that is gone. How do I coupe with that? How can I move on and be happy after knowing all the hard work we put in and all the love is out the door now. I just can't wrap my head around knowing that the girl I thought I would marry will now move on and live her life away from me till the day we both die.

I still do care for her, yes I still love her even after everything that happened at the end. I feel so alone and lost even though the actual break up was months ago, I'm honestly just afraid of the future now.

- **Base**: You need to take a break from your ex and focus on yourself. Figure out who you are and what you want out of life without her. Surround yourself with people who care about you. Seek therapy to help you cope with your feelings and learn healthy ways to manage them. Utilize positive activities, such as connecting with family, friends and hobbies. Remind yourself that everything is never going to be perfect and you have the strength to move on and lead a happy life despite what happened in the past.

- **SFT** : I love my ex and can't cope with the fact that she is 100% done with me. How do I move ahead and be happy? How am I to feel after being in love with my ex for a year. So lost and what do I do?

- R^{RM}: Despite the breakup months ago, I'm still hurting and scared of the future. I can't forget how much we had together and how happy we were, yet I know I have to move on and learn to live without her. I care for her and love her, but I'm afraid of the unknown of what lies ahead.

- $R_{\beta*}$ I'm still deeply hurt and angry about the breakup, and I'm scared I'll never find that same connection with someone else. I don't know how to cope with the fact that the girl I thought I would marry is now moving on without me, and I feel lost and alone without her. I still care about her and I love her, but

I'm afraid of the direction my life has taken now.

- **RCfD** I cannot accept the fact that my ex and I will never be together again. We had so many plans and built so much love, only to have it all end suddenly. Despite the heartbreak, I still care for her and love her, and it hurts that she will move on and live her life away from me. I feel lost and afraid of the future, not knowing when or if she'll ever come back.
- R^{RM} **&** R^{length} I can't accept the fact that my ex and I will never be together again and it's difficult to move on. I care for her and fear the future.
- **RCfD with length regularization**: The narrator is struggling to accept the breakup of their relationship and is filled with sadness and uncertainty about the future. They still care for and love their ex-partner, but are afraid of what the future holds without them.

Prompt 2: Summarize: This is my first post so please be kind :)

I know that lots of people often feel confused when they come out of a long-term relationship. They think they have forgotten how to be single, or how to flirt/date.

I am one of these people.

The problem is, my relationship started when I had just turned 16. I have never been single - as an adult. That might sound silly. But the only time I have ever flirted or dated was as an over-confident, hormone-riddled teenager.

Now I have a pretty demanding job, responsibilities blah blah... And I just don't know how to this!

I'm no way in a rush to get into a new relationship, but that doesn't mean I want to be completely alone in the mean time.

If anyone has experienced anything similar, or just generally has some advice, it would be greatly appreciated!

- **Base**: It can be a challenge to transition from being in a relationship to being single again, but there are a few things you can do to ease this process.

  Firstly, it is important to give yourself time and space to adjust. Avoid rushing into a new relationship, but try to also avoid becoming too introverted. Engage in activities and hobbies that you enjoy, and use this time to meet new people.

  It also helps to keep busy with stress-relieving activities, such as exercise, art, or writing, to help manage any feelings of anxiety or doubt.
- **SFT** : I am feeling a spike in my sex drive post exercise and want to know if having sex while being physically active is related to having an increased sex drive. What are your experiences? Have you ever had similar experiences? What do you think would explain my experience? Appreciate your help.
- R^{RM}: I'm writing my first post ever and am seeking advice on how to date and flirt after being in a long-term relationship for many years. I'm feeling confused and unsure of how to navigate the dating world as an adult since I've never really done it before. I'm scared of making mistakes and don't know where to start. Any advice would be appreciated!
- $R_{\beta*}$ I am at a point in my life where I have never been single and got out of a long-term relationship. I'm feeling confused and don't know how to flirt/date, as well as having responsibilities. I'm just looking for some advice from those who may have been in a similar situation before.
- RCfD My first post on the subject was about the difficulties posed by being in an intense relationship for a long time and then, all of a sudden, finding oneself single and not knowing how to act or behave. My relationship started when I was 16 and I've never been single as an adult - not even for a few months. This makes it difficult to flirt and date, as I have virtually no experience in this area. I'm not looking for a relationship at the moment, but I would like to enjoy the company of others and learn how to do so, as I feel I'm missing out on something. Any advice from those who have experienced a similar situation would be appreciated.
- R^{RM} & R^{length} I'm a 23-year-old woman who is looking for guidance on how to date and flirt after being in a long-term relationship for 6 years. Any advice is appreciated!

- RCfD with length regularization: The writer is feeling uncertain about how to date and flirt now that they are single after being in a long-term relationship. They would appreciate advice from those who have been in a similar situation.

Prompt 3: Summarize: On Sunday, I ran my first 10-mile race. While this is a great accomplishment I have noticed a significant spike in my sex drive during the week leading up to the race and in the days since I ran; about 10 days in total. I workout 2-3 times per week for about 45 minutes each time and have average eating habits. During the week leading up to the race, I ate mostly a plant-based diet - salad for lunch and dinner, fish for protein, etc. - and worked out 3 times and ran only twice (Not great prep, I know.) This helped me lose about 2 or 3 pounds. Now, I cannot stop thinking about sex! This morning (Thursday) I spent about 3 hours watching porn and "distracted" myself by going to a midday fitness class. Now I want to jerk off again or have find someone to sex. I don't know what to do. This is both embarrassing and confusing. I want to continue working out but am honestly afraid being fit will turn me into a sex addict. Background: I played soccer in college and recall a similar correlation during my off season workouts. I am now in my late 20s and no longer train 5-6 days a week as was the norm in college. This is the most physically active I have been since my early 20s. Question: Is there research that has connected a higher sex drive (thinking more about sex) with a increased physical activity (working out)? Also, share your experiences because I feel like an outlier here.

- **Base**: It is definitely not unprecedented for an increase in physical activity to lead to an increase in sex drive; in fact, there is evidence to support this relationship. For example, one study found that individuals who exercised at a moderate level had substantially higher levels of sexual desire than those who didn't exercise. Additionally, research has found that exercise can lead to positive changes in mood, self-esteem and body image, which can, in turn, lead to an increase in confidence and sexual desire.

  For reference, when you exercised regularly and intensely as part of your fitness routine, did you notice a change in your sex drive?

- **SFT** : I cannot stop thinking about sex, is there research that connects increased physical activity with an increased sex drive? I am a 27 yr old sports young man. Please share your experiences. Shine some light on this subject. I need insight before it takes over my life.

- R^{RM}: I recently ran my first 10-mile race and since then have found myself suddenly extremely sexually aroused. This is the first time this has ever happened to me and I'm quite confused and concerned. I'm working out more regularly than usual and lost a few pounds in the process, but I haven't changed my diet or workout routine much. I'm worried that this is a sign of me developing a sex addiction and am seeking advice from the community as to whether there is any scientific research on this correlation and if anyone has experienced a similar situation.

- $R_{\beta *}$ There has been very limited research on the link between physical activity and increased libido. However, there was one study in 2008 that found that recreational endurance athletes reported higher levels of sexual desire and satisfaction than non-athletes. So, it is possible that increased physical activity can lead to an increase in sex drive. Other than this study though, there is not much more research on the link between physical activity and increased libido.

  As for my own experiences, I can attest to anecdotally noticing that when I increase my physical activity, my sex drive tends to increase as well. From my own observations though, I have noticed that the increase in sex drive is generally mild and not overly extreme. I have also found that the increase in sex drive tends to reduce after my workouts become more regular and my physical activity becomes more intense. It seems that my body and mind need a balance between physical activity and rest

- **RCfD** In the days leading up to running my first 10-mile race and in the days since, I have noticed a significant spike in my sex drive. I am averaging about 10 days of thinking about sex. This has been unexpected and quite confusing. I workout 2-3 times per week for about 45 minutes each time and have average eating habits. During the week leading up to the race, I ate mostly a plant-based diet - salad for lunch and dinner, fish for protein, etc. - and worked out 3 times and ran only twice (Not great prep, I know.). This helped me lose about 2 or 3 pounds. Now, I cannot stop thinking about sex! This morning (Thursday) I spent about 3 hours watching porn and "distracted" myself by going to a midday fitness class. Now I want to jerk off again or

- R^{RM} & R^{length} I recently ran my first 10-mile race and have noticed a drastic increase in my sex drive. I have not been this physically active in several years and am concerned that I may be turning into a sex addict. Any research or experiences to share?
- RCfD with length regularization: The speaker is inquiring about whether there is research linking an increased sex drive with increased physical activity and asked for input from others who have experienced a similar phenomenon.