On Reinforcement Learning and Distribution Matching for Fine-Tuning Language Models with no Catastrophic Forgetting

Tomasz Korbak*
University of Sussex
tomasz.korbak@gmail.com

Hady Elsahar
Naver Labs Europe
hady.elsahar@gmail.com

Germán Kruszewski
Naver Labs Europe
german.kruszewski@naverlabs.com

Marc Dymetman†
marc.dymetman@gmail.com

Abstract

The availability of large pre-trained models is changing the landscape of Machine Learning research and practice, moving from a "training from scratch" to a "fine-tuning" paradigm. While in some applications the goal is to "nudge" the pre-trained distribution towards preferred outputs, in others it is to steer it towards a different distribution over the sample space. Two main paradigms have emerged to tackle this challenge: Reward Maximization (RM) and, more recently, Distribution Matching (DM). RM applies standard Reinforcement Learning (RL) techniques, such as Policy Gradients, to gradually increase the reward signal. DM prescribes to first make explicit the target distribution that the model is fine-tuned to approximate. Here we explore the theoretical connections between the two paradigms, and show that methods such as KL-control developed for RM can also be construed as belonging to DM. We further observe that while DM differs from RM, it can suffer from similar training difficulties, such as high gradient variance. We leverage connections between the two paradigms to import the concept of *baseline* into DM methods. We empirically validate the benefits of adding a baseline on an array of controllable language generation tasks such as constraining topic, sentiment, and gender distributions in texts sampled from a language model. We observe superior performance in terms of constraint satisfaction, stability and sample efficiency.

1 Introduction

Pre-trained language models (Devlin et al., 2019; Radford et al., 2019) are changing the landscape of Machine Learning research and practice. Due to their strong generative capabilities many studies have found it sufficient to "nudge" these models to conform to global preferences defined over the generated sequences instead of training from scratch using annotated data. These preferences could include topic and sentiment (Dathathri et al., 2020), valid musical notes and molecular structures (Jaques et al., 2017a), code compilability (Korbak et al., 2021), reducing distributional biases (Khalifa et al., 2021; Weidinger et al., 2021), evaluation metrics for Machine Translation and Summarization (Ranzato et al., 2016; Bahdanau et al., 2016), or direct human feedback (Ziegler et al., 2019; Stiennon et al., 2020). This large body of studies is driven by two different paradigms: *Reward Maximization* (RM) and *Distribution Matching* (DM).

*Work partly done during an internship at Naver Labs Europe.
†Independent Researcher. Work done at Naver Labs Europe.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

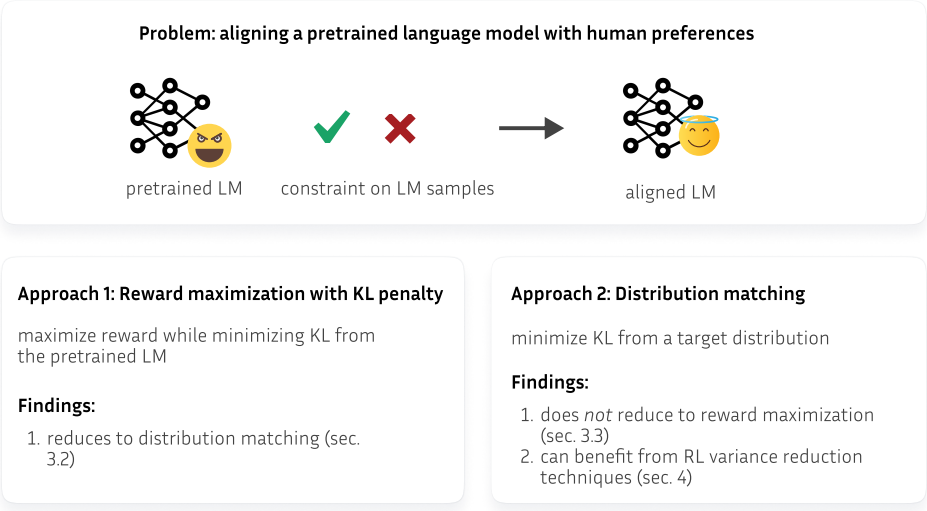


Figure 1: In this study we make a connection between two popular paradigms for aligning language models to human preferences: Reward maximization (RM) and Distribution matching (DM).

Reward Maximization RM intuitively nudges pre-trained models towards certain preferences by providing global sequence-level rewards when the model generates outputs that satisfy desired features. For instance, if the model is producing toxic content, we can apply Reinforcement Learning (RL) techniques to discourage it from producing similar content. However, naively applying RL yields a model that can undergo *catastrophic forgetting* of its original distribution. For example, it can degenerate into producing a single nonsensical but at least nontoxic sequence. Although several studies have considered hand-crafting general rewards to ensure desirable features like fluency (Liu et al., 2016a; Tambwekar et al., 2019), coming up with complete or perfect rewards is highly non-trivial (Wu et al., 2016; Vedantam et al., 2015). This has sparked a wide discussion on the overall effectiveness of RM for some tasks such as machine translation (Choshen et al., 2020; Kiegeland & Kreutzer, 2021).

Reward Maximization with KL-Control To tackle the aforementioned issues of "catastrophic forgetting", several studies, still under an RM paradigm, have considered incorporating a distributional term inside the reward to be maximized. In particular Jaques et al. (2017b, 2019) and Ziegler et al. (2019) or more recently Stiennon et al. (2020), Ouyang et al. (2022), Bai et al. (2022) and Perez et al. (2022) have applied variations of KL-control (Todorov, 2007; Kappen et al., 2012) which adds a penalty term to the reward term so that the resulting policy does not deviate too much from the original one in terms of KL-divergence. The overall objective with the KL-penalty is maximized using an RL algorithm of choice including: PPO (Schulman et al., 2017a) as in Ziegler et al. (2019) or Bai et al. (2022) or Q-learning (Mnih et al., 2013) as in Jaques et al. (2017b). Adding this *distributional* KL-penalty to the reward raises some important questions: What effect does it have on the shape of the optimal policy? Does this new objective have any interpretation from a distributional perspective?

Distribution Matching A different recent paradigm for fine-tuning language models to satisfy downstream preferences formulates the problem as Distribution Matching (DM). This paradigm consists of two steps: first a target distribution incorporating the desired preferences is defined as an Energy-Based Model (LeCun et al., 2006). Then the forward KL divergence is minimized between this target distribution and an auto-regressive policy using a family of algorithms referred to as Distributional Policy Gradients (DPG) (Parshakova et al., 2019b; Khalifa et al., 2021; Korbak et al., 2021, 2022a). This approach capitalizes on the flexibility of EBMs in specifying the target distribution. For example, the EBM can be defined so that it conforms to all downstream preferences while its corresponding normalized distribution has a minimal KL divergence from the original, pre-trained language model, therefore tackling the problem of "catastrophic forgetting" (Khalifa et al., 2021). Interestingly, this DM paradigm can also deal with *distributional* preferences, for instance, for de-biasing language models by specifying that the generated sequences should be gender-balanced,

i.e. that 50% of generations contain female mentions. Such distributional constraints cannot be defined in the RM paradigm where a reward is calculated for a single sequence.

We can notice the promises and limitations of these two paradigms for fine-tuning language models. RM approaches are equipped with an arsenal of RL algorithms and optimization techniques that can be efficient in reward maximization, however they lack the distributional aspect to avoid catastrophic forgetting and impose distributional preferences over LMs. DM approaches are suited to tackle those limitations, however, the family of DPG algorithms currently used is not as rich as its RL counterpart.

While the connections between these two seemingly distinct paradigms have been noted (Parshakova et al., 2019b; Korbak et al., 2022b), they have not been explored in detail. Clarifying such connections might help import ideas from one approach to the other. This is our goal in this paper, detailing the nuanced connections and applying them to a case-study in variance reduction. Overall, our contributions are the following:

- We clarify relations between the RM and DM paradigms through a detailed comparison between the family of DPG algorithms and Policy Gradients (Table 1), stressing the differences between *parametric* and *non-parametric* rewards that are important in this regard.
- We introduce an interpretation of KL-control techniques from a distribution matching perspective, placing such techniques at an intermediate place between RM and DM (Theorem 1).
- We show how these connections can enable cross-pollination between the two perspectives by applying *baselines* — a variance reduction technique from RL — to DPG and derive a particular choice of a baseline (Facts 1 and 2). On an array of controllable language generation experiments, we show that adding baselines leads to superior performance on constraint satisfaction (Figure 3), stability on small batch sizes, and sample efficiency (Figure 4).

2 Background

Standard Policy Gradients One popular method for adapting the behaviour of language models to certain preferences has been that of assigning a "reward" score $R(x)$ for sequences x sampled from an autoregressive language model (policy) π_θ. Then, the simplest policy gradient algorithm in reinforcement learning, namely, REINFORCE (Williams, 1992a), aims to find the policy $\pi_\theta(x)$ that maximizes the average reward $\mathbb{E}_{x\sim\pi_\theta}R(x)$, and this leads, via the so-called "log derivative trick", to a gradient ascent algorithm that iteratively samples x from π_θ and update parameters by increments proportional to $R(x)\nabla_\theta \log \pi_\theta(x)$ via the following identity:

$$\nabla_\theta \mathbb{E}_{x\sim\pi_\theta}R(x) = \mathbb{E}_{x\sim\pi_\theta}R(x)\nabla_\theta \log \pi_\theta(x). \tag{1}$$

KL-control (Todorov, 2007; Kappen et al., 2012), was leveraged by Jaques et al. (2017b, 2019) and Ziegler et al. (2019) to include a KL penalty term in the reward function to penalize large deviations from the original pretrained model $a(x)$, weighted by a free hyperparameter β to control the trade-off between the two goals. That is, they maximize the expectation $\mathbb{E}_{x\sim\pi_\theta}R^z_\theta(x)$, where:

$$R^z_\theta(x) \doteq r(x) - \beta \log \frac{\pi_\theta(x)}{a(x)}. \tag{2}$$

Distributional Policy Gradients (DPG) (Parshakova et al., 2019b) is a recent approach used to fit an autoregressive policy π_θ to the distribution $p(x) = P(x)/Z$ induced by the EBM $P(x)$, where $Z = \sum_x P(x)$ is the normalization constant (partition function). Given an arbitrary EBM $P(x)$, DPG optimizes the loss function $D_{\mathrm{KL}}(p, \pi_\theta)$ with respect to the parameters θ of an autoregressive model π_θ, a loss which is minimized for $\pi_\theta = p$. The KL-divergence minimization objective leads to a gradient estimate of the form:

$$\nabla_\theta D_{\mathrm{KL}}(p, \pi_\theta) = -\nabla_\theta \mathbb{E}_{x\sim p} \log \pi_\theta(x) \tag{3}$$

$$= -\sum_x p(x)\nabla_\theta \log \pi_\theta(x) = -\frac{1}{Z}\sum_x P(x)\nabla_\theta \log \pi_\theta(x) \tag{4}$$

$$= -\frac{1}{Z}\ \mathbb{E}_{x\sim\pi_\theta}\frac{P(x)}{\pi_\theta(x)}\nabla_\theta \log \pi_\theta(x). \tag{5}$$

3 Reward Maximization vs Distribution Matching

In the previous section, we have summarized three approaches that have been suggested for fine-tuning language models. Two of them can be characterized as "Reward Maximization" (RM): Standard Policy Gradients (PG) and KL-control. On the other hand, DPG clearly belongs to the realm of "Distribution Matching" (DM) as it first defines the target distribution and then optimizes a policy to match it. In the rest of this section, we will explore connections between these two seemingly distinct concepts and, in the following section, we will exploit them to improve DM-based methods.

3.1 Standard vs. Parametric Rewards

Let us start with distinguishing between a "parametric reward" R_θ which depends on θ and a standard reward R, which does not. If we wished to maximize the expected parametric reward, $\mathbb{E}_{\pi_\theta} R_\theta(x)$, we would follow its gradient, leading to the identities:

$$\nabla_\theta \mathbb{E}_{x\sim\pi_\theta} R_\theta(x) = \nabla_\theta \sum_x \pi_\theta(x) R_\theta(x) = \sum_x \pi_\theta(x) \nabla_\theta R_\theta(x) + \sum_x R_\theta(x) \nabla_\theta \pi_\theta(x) \tag{6}$$

$$= \sum_x \pi_\theta(x) \nabla_\theta R_\theta(x) + \sum_x \pi_\theta(x) R_\theta(x) \nabla_\theta \log \pi_\theta(x) \tag{7}$$

$$= \underbrace{\mathbb{E}_{x\sim\pi_\theta} \nabla_\theta R_\theta(x)}_{\text{RG-term}} + \underbrace{\mathbb{E}_{x\sim\pi_\theta} R_\theta(x) \nabla_\theta \log \pi_\theta(x)}_{\text{PG-term}}. \tag{8}$$

Equation (8) is the sum of two terms: the first one, the "RG-term" (Reward Gradient term), involves the gradient of the reward. The second one, the "PG-term" (Policy Gradient term), was obtained using the "log derivative trick" and involves the gradient of the policy *stricto sensu*. In standard RL, where the reward does *not* depend on θ, the RG-term disappears and the gradient of expected reward consists solely of the PG-term. However, when R_θ depends on θ, the gradients are distinct (apart from specific cases where the RG-term evaluates to 0, as we will see below).

3.2 KL-control as Distribution Matching

Adding a KL-penalty term to the reward (as in the case of KL-control) leads to a parametric reward. However, due to the particular form of its objective, the RG-term actually *vanishes*,[3] leaving only the PG-term $\mathbb{E}_{x\sim\pi_\theta} R^z_\theta(x) \nabla_\theta \log \pi_\theta(x)$ and simplifying the tuning procedure to a standard Policy Gradient. While this algorithm falls under the RM paradigm, here we argue that is its nature is multifaceted, and explore deeper connections with the DM paradigm. More precisely, the maximization of reward with the KL penalty term is equivalent to a distributional matching with an underlying emergent sequential EBM, a remark that already reveals some similarities with DPG.[4]

Theorem 1. *Consider the following EBM:*

$$P_z(x) = a(x) e^{r(x)/\beta} \tag{9}$$

and let p_z be the normalized distribution $p_z(x) = \frac{1}{Z} P_z(x)$, with $Z = \sum_x P_z(x)$. Then:

(i) $\arg\max_{\pi_\theta} \mathbb{E}_{x\sim\pi_\theta} R^z_\theta(x) = \arg\min_{\pi_\theta} D_{\mathrm{KL}}(\pi_\theta, p_z)$*;*

(ii) $\arg\max_{\pi\in\mathcal{D}(X)} \mathbb{E}_{x\sim\pi} R^z_\pi(x) = p_z$*, where $\mathcal{D}(X)$ is the family of all distributions over X, and $R^z_\pi(x) \doteq r(x) - \beta \log \frac{\pi(x)}{a(x)}$.*

Proof. A simple way to prove this is to notice that the expectation of the reward R^z_θ has a monotonically decreasing relationship with the *reverse* KL divergence between π_θ and p_z:

$$D_{\mathrm{KL}}(\pi_\theta, p_z) = \mathbb{E}_{x\sim\pi_\theta} \log \frac{\pi_\theta(x)}{p_z(x)} = \mathbb{E}_{x\sim\pi_\theta} \Big[\log \pi_\theta(x) - \log \frac{1}{Z} a(x) e^{r(x)/\beta} \Big]$$

[3] This is because $\mathbb{E}_{\pi_\theta} \nabla_\theta R^z_\theta(x) = -\beta\, \mathbb{E}_{\pi_\theta} \nabla_\theta \log \pi_\theta(x) = 0$, via the identity $\mathbb{E}_{\pi_\theta} \nabla_\theta \log \pi_\theta(x) = \sum_x \pi_\theta(x) \nabla_\theta \log \pi_\theta(x) = \sum_x \nabla_\theta \pi_\theta(x) = \nabla_\theta \sum_x \pi_\theta(x) = 0$.

[4] The optimal policy p_z is briefly mentioned in (Ziegler et al., 2019) without reference or derivation. The proof, which reveals a connection to the reverse KL divergence from π_θ, is ours.

	Policy Gradients	**DPG**
Reward	$R(x)$	$R_\theta(x) = \frac{P(x)}{\pi_\theta(x)}$
∇_θ	$\mathbb{E}_{x\sim\pi_\theta} R(x) \nabla_\theta \log \pi_\theta(x)$	$\mathbb{E}_{x\sim\pi_\theta} \frac{P(x)}{\pi_\theta(x)} \nabla_\theta \log \pi_\theta(x)$
Baseline	$\mathbb{E}_{x\sim\pi_\theta} R(x)$	Z
∇_θ **with Baseline**	$\mathbb{E}_{x\sim\pi_\theta} \Big[R(x) - \mathbb{E}_{x\sim\pi_\theta} R(x)\Big] \nabla_\theta \log \pi_\theta(x)$	$\mathbb{E}_{x\sim\pi_\theta} \Big[\frac{P(x)}{\pi_\theta(x)} - Z\Big] \nabla_\theta \log \pi_\theta(x)$

Table 1: A comparison between Policy Gradients (Sutton et al., 1999) and Distributional Policy Gradients (Parshakova et al., 2019b) forms of Reward, Baseline, and Gradient of the loss function (the PG-term) before (∇_θ) and after (∇_θ with Baseline) including a baseline for variance reduction .

$$= \log Z - \frac{1}{\beta} \mathbb{E}_{x\sim\pi_\theta} \Big[r(x) - \beta \log \frac{\pi_\theta(x)}{a(x)}\Big] = \log Z - \frac{1}{\beta} \mathbb{E}_{x\sim\pi_\theta} R^z_\theta(x), \tag{10}$$

so that the $\arg\min_{\pi_\theta} D_{\mathrm{KL}}(\pi_\theta, p_z)$ coincides with the $\arg\max_{\pi_\theta} \mathbb{E}_{x\sim\pi_\theta} R^z_\theta(x)$, proving (i). On the other hand, $\arg\min_{\pi\in\mathcal{D}(X)} D_{\mathrm{KL}}(\pi, p_z)$, which also corresponds to $\arg\max_{\pi\in\mathcal{D}(X)} \mathbb{E}_{x\sim\pi} R^z_\pi$ because of (i) applied to a family $\pi_{\theta'}$ covering $\mathcal{D}(X)$ in full, is just p_z, concluding the proof. □

Overall, we can conclude that the addition of the distributional term (KL-penalty) to the reward does indeed provide a DM interpretation, namely in terms of minimizing the reverse KL divergence with an emergent underlying distribution $p_z(x)$. We note that $p_z(x)$ does not correspond to a free and explicit choice of EBM (e.g. one that balances the gender and topic distributions of a language model). Instead equation (9) appears in a restrictive format, which is implicitly defined by the reward R^z_θ, along with a β hyperparameter without a clear meaning. By contrast, the DPG algorithms are designed to perform DM on any EBM specification, corresponding to an explicit distributional objective.

3.3 Similarities and Differences between DPG and Policy Gradients

In the previous subsection, we have connected KL-control, a method designed under a RM paradigm, to DM. Now, we turn to the converse question of whether DPG, a DM method, can be connected to RM. We begin by noting that after defining $R_\theta = \frac{P(x)}{\pi_\theta(x)}$, the DPG gradient $\mathbb{E}_{x\sim\pi_\theta} \frac{P(x)}{\pi_\theta(x)} \nabla_\theta \log \pi_\theta(x)$ acquires the format of the PG-term $\mathbb{E}_{\pi_\theta} R_\theta \nabla_\theta \log \pi_\theta(x)$.

However, the DM objective of DPG *cannot* be considered as maximizing the average "reward" $R_\theta(x) = \frac{P(x)}{\pi_\theta(x)}$, as this would require adding also the RG-term $\mathbb{E}_{\pi_\theta} \nabla_\theta \frac{P(x)}{\pi_\theta(x)}$ into the gradient, which in this case does not vanish.

Nonetheless, the analogy behind this gradient term is more fruitful than it first appears. As a matter of fact, DPG gradient estimates suffer from the same high-variance problems as with standard PG. While the objective of DPG (distribution matching) is different from that of Policy Gradients (reward maximization), DPG also needs to estimate the PG-term $\mathbb{E}_{\pi_\theta} R_\theta(x) \nabla_\theta \log \pi_\theta(x)$ at a *given* value of θ, using a batch of samples x. For such a *fixed* θ, we can define provisionally set $R(x) \doteq R_\theta$ and the problem of gradient estimation *for this fixed* θ is identical to the estimation $\mathbb{E}_{x\sim\pi_\theta} R(x) \nabla_\theta \log \pi_\theta(x)$ based on a set of samples x in standard RL. Therefore, techniques that have been developed to reduce the variance of the gradients estimates in RL can be ported to DPG insofar as we are computing the gradient estimates *at a given* θ. In Section 4, we show how one can import one such variance reduction technique to the DPG: baselines.

4 A Case Study on Variance Reduction

Baselines are a standard variance reduction technique in the context of Policy Gradients (Sutton & Barto, 2018). The idea is to subtract from the reward $R(x)$ a value B that does not introduce bias to the gradients but may change variance. After the introduction of baseline, equation (1) then takes the following form:

$$\nabla_\theta \mathbb{E}_{\pi_\theta} R(x) = \mathbb{E}_{\pi_\theta} [R(x) - B] \nabla_\theta \log \pi_\theta(x). \tag{11}$$

In standard RL, the simplest form of baseline B is just the average of the rewards for the policy:[5]

$$B^{\text{RL}} = \mathbb{E}_{x\sim\pi_\theta} R(x). \tag{12}$$

Following the same methodology of taking the baseline to be the expectation of the reward term, we can obtain a remarkably simple form of a baseline for DPG: [6]

$$B = \mathbb{E}_{x\sim\pi_\theta} \frac{P(x)}{\pi_\theta(x)} = \sum_x \pi_\theta(x) \frac{P(x)}{\pi_\theta(x)} = \sum_x P(x) = Z. \tag{13}$$

Fact 1. *Subtracting B from $R_\theta(x)$ does not introduce bias into DPG gradient estimates.*

Proof. Let us rewrite the DPG gradient in (5) with the added baseline $B = Z$:

$$\begin{aligned}\mathbb{E}_{x\sim\pi_\theta}\Big[R_\theta(x) - Z\Big]\nabla_\theta \log \pi_\theta(x) &= \mathbb{E}_{x\sim\pi_\theta} R_\theta(x)\nabla_\theta \log \pi_\theta(x) - Z\,\mathbb{E}_{x\sim\pi_\theta}\nabla_\theta \log \pi_\theta(x) \\ &= \mathbb{E}_{x\sim\pi_\theta} R_\theta(x)\nabla_\theta \log \pi_\theta(x) - Z\Big[\sum_x \nabla_\theta \pi_\theta(x)\Big]\end{aligned} \tag{14}$$

Here, the second term does not introduce bias because $Z\Big[\sum_x \nabla_\theta \pi_\theta(x)\Big] = 0$, leaving us with the exact same form of gradient as in the original DPG algorithm. □

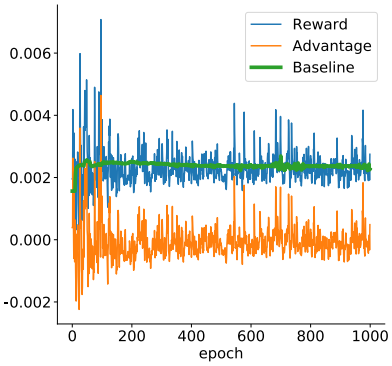


Figure 2: Values of reward, advantage and the baseline for first 1000 epochs of a point-wise constraint experiment.

Note that since B^{RL} depends on θ, it has to be be re-estimated after each gradient update. On the other hand, B does *not* depend on θ, which is an advantage because B could be now estimated by averaging over samples from *all* the different θ's without introducing bias, leading to a more accurate estimation. See Table 1 for a comparison of these two forms of baselines.

Algorithm 1 KL-Adaptive DPG with baseline

Require: P, initial generative model a
$\pi_\theta \leftarrow a$, $q \leftarrow a$
for each iteration **do**
  for each episode **do**
    sample x from $q(\cdot)$
    $\theta \leftarrow \theta + \alpha^{(\theta)}\left[\frac{P(x)}{q(x)} - Z\frac{\pi_\theta(x)}{q(x)}\right]\nabla_\theta \log \pi_\theta(x)$
  if $D_{\text{KL}}(p||\pi_\theta) < D_{\text{KL}}(p||q)$ **then**
    $q \leftarrow \pi_\theta$
Ensure: π_θ

The off-policy DPG version introduced in (Parshakova et al., 2019b) and its KL-adaptive variant (Khalifa et al., 2021) sample a proposal distribution q instead of the policy π_θ. Then, the baseline takes the form

$$B^{\text{off}}(x) = Z\frac{\pi_\theta(x)}{q(x)}, \tag{15}$$

where the $\frac{\pi_\theta(x)}{q(x)}$ term is an importance weight correcting for the bias introduced by sampling from q. Similarly to the DPG case, we can prove the following (see Appendix C):

Fact 2. *Subtracting $B^{off}(x)$ from $R_\theta(x)$ does not bias the off-policy DPG gradient estimates.*

In practice, as shown on Figure 2, adding a baseline to KL-adaptive DPG (Algorithm 1) centers the advantage (defined as $A \doteq \frac{P(x)}{q(x)} - Z\frac{\pi_\theta(x)}{q(x)}$) around 0 leading to better performance on: convergence (section 4.3), stability on small batch sizes (section 4.4), and variance reduction (section 4.5).

[5] While this baseline is not optimal (proof Appendix C.1), it is widely used in practice.

[6] In the scope of this paper, our focus is on importing to DPG simple constant baselines. The advantage is that this is a technique that is not impacted by the fact that R_θ depends on θ: it can be applied "θ-locally" to provide a more accurate estimate of $\mathbb{E}_{x\sim\pi_\theta} R_\theta(x)\nabla_\theta \log \pi_\theta(x)$ for a *fixed* θ, irrespective of the values of $R_{\theta'}$ elsewhere, while variance reduction techniques that involve several $\theta's$ simultaneously raise additional challenges for parametric rewards.

4.1 Generation with Distributional Control

We investigate the benefits of adding a baseline to the DPG algorithm, on the Generation with Distributional Control (GDC) (Khalifa et al., 2021) framework. GDC makes use of DPG to control the properties of pre-trained language models to satisfy certain constraints. In our experiments, follow target distribution form of Parshakova et al. (2019a), Khalifa et al. (2021) and Korbak et al. (2022a), in which the EBM $P(x)$ is defined so that its normalized variant $p(x)$ matches a set of desired moments constraints on given features $\phi_i(x)$, while having a minimal KL divergence $D_{\mathrm{KL}}(p, a)$ from an original pretrained language model a, to avoid catastrophic forgetting.

These constraints are expressed as conditions $\bar{\mu}_i = \mathbb{E}_{x\sim p}\phi_i(x)$, for $i \in \{1, \ldots, n\}$, by which the moments (expectations) under the distribution p of each feature $\phi_i(x)$ are required to take certain desired values $\bar{\mu}_i$. For instance, let $\phi_1(x) = 1$ iff the topic of x is science and $\phi_2(x) = 1$ iff x mentions a female person, then imposing moments $\bar{\mu}_1 = 1$ and $\bar{\mu}_2 = 0.5$ constrains the language model p to only generate sequences about science, half of which mention females. $P(x)$ is uniquely determined by the following form:[7]

$$P(x) = a(x)e^{\sum_{i=1}^{n} \lambda_i \phi_i(x)}, \tag{16}$$

where λ_i terms control the moments μ_i of the associated features, which can be estimated through self-normalized importance sampling (Owen, 2013); and then, to make the moments match the desired values, the λ_i terms can be optimized through SGD (Parshakova et al., 2019a).

4.2 Experimental setup

We evaluate our method on an array of 10 controlled text generation tasks. For each, given a pre-trained language model $a(x)$, and a set of constraints, the objective of each fine-tuning method is to obtain a fine-tuned language model π_θ that satisfies the imposed constraints while deviating as minimally as possible from the original language model $a(x)$.

Constraints are defined as a set of binary features $\{\phi_i\}$ and their corresponding desired percentages (moments) $\{\bar{\mu}_i\}$ within the generations of the target language model. Based on the value of the moment constraints these 10 tasks are divided into 6 tasks of pointwise constraints (for which $\bar{\mu}_i = 1$), 2 tasks of distributional constraints ($0 < \bar{\mu}_i < 1$) and 2 tasks of mixed type constraints (hybrid):

(a) Single-word constraints, where $\phi(x) = 1$ iff the a given word appears in the sequence x. We experiment with frequent words (task 1: "amazing", original frequency: 10^{-4}) and (task 2: "WikiLeaks", original frequency: 10^{-5}) rare words,

(b) Wordlist constraints, where $\phi(x) = 1$ iff x contains at least one word from a given list. We consider lists of word associated with politics (task 3) and science (task 4) published by Dathathri et al. (2020),

(c) Sentiment classifier constraints, where $\phi(x) = 1$ if x is classified as positive (task 5), or negative (task 6) by a pre-trained classifier published by Dathathri et al. (2020).

(d) A single distributional constraint where $\phi(x) = 1$ iff x contains a female figure mention, and $\bar{\mu} = 0.5$ (task 8),

(e) A set of four distributional constraints: $\phi_i(x) = 1$ iff x contains at least one of the words in the "science", "art", "sports" and "business" wordlists (compiled by Dathathri et al. (2020)), respectively. For each i, $\bar{\mu}_i = 0.25$ (task 8),

(f) Hybrid constraints where $\phi_1(x) = 1$ iff x contains more female than male pronouns, $\bar{\mu}_1 = 0.5$ and $\phi_2(x) = 1$ iff x contains at least one of the words from the "sports" wordlist (task 9) or "politics" wordlist, $\bar{\mu}_2(x) = 1$ (task 10).

Methods We modify the GDC framework Khalifa et al. (2021), namely its KL-DPG algorithm, to include a baseline as shown in Algorithm 1. We refer to this method as **GDC++**. In addition to comparing **GDC++** with **GDC** we compare with two reward maximization baselines: **Reinforce** (Williams, 1992b) and **Ziegler** (Ziegler et al., 2019). Reinforce tries to maximize the expected reward $\mathbb{E}_{x\sim\pi_\theta} R(x)$, where $R(x) = 1$ if and only if the pointwise constraints are met. Ziegler instantiates the KL-control approach: its objective includes a KL penalty term for departures from a. Following (Khalifa et al., 2021), for hybrid and distributional constraints (tasks 8-10) we compare

[7] For a more precise formulation of this EBM, see (Khalifa et al., 2021).

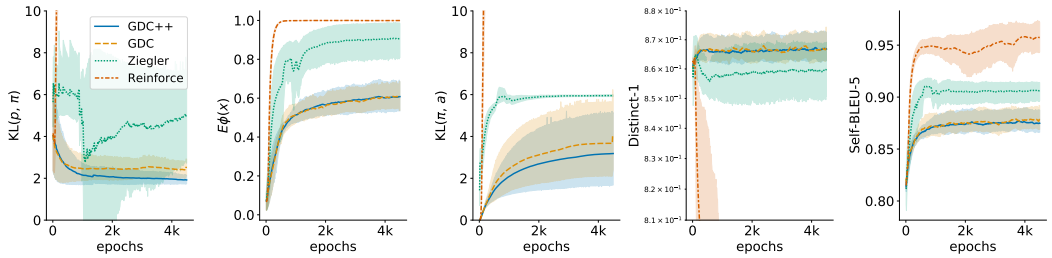


Figure 3: Evaluation metrics: $D_{\mathrm{KL}}(p, \pi_\theta)$ (↓ better), $\mathbb{E}_{\pi_\theta}\phi(x)$ (↑ better), $D_{\mathrm{KL}}(\pi_\theta, a)$ (↓ better), Self-BLEU-5 (↓ better), and Distinct-1 (↑ better) aggregated over 6 pointwise constraints experiments (tasks 1-6) for policies obtained from GDC++, GDC, Ziegler and Reinforce. See Figure 6 for aggregated distributional constraints experiments. In the Appendix Figures 7-10 and Table 4 contain individual view and final results of each run.

only GDC and GDC++ because the RM objective of Ziegler and Reinforce is not equipped to handle them.

Metrics We report the following metrics at each validation step over batches of samples from π_θ:

1. $\mathbb{E}_{x\sim\pi_\theta}\phi_i(x)$, measuring the ability to reach the target moment of the i-th feature.
2. $D_{\mathrm{KL}}(p, \pi_\theta)$, the forward KL divergence from the optimal target distribution p,[8]
3. $D_{\mathrm{KL}}(\pi_\theta, a)$, the reverse KL divergence from the original pretrained language model a.
4. Distinct-n score, a measure of text diversity in terms of the frequency of repetitions within a single sample x, proposed by (Li et al., 2016a).
5. Self-BLEU-n, a measure of text diversity on a distributional level *across* samples proposed by (Zhu et al., 2018), ensuring that policies don't converge into limited number of sequences that satisfy the imposed constraints Caccia et al. (2020).

Training details For tasks 1-6, we use a pre-trained GPT-2 small with 117M parameters (Radford et al., 2019) as the original language model a. For tasks 7-10, a is the same pre-trained model additionally fine-tuned on the WikiBio (Lebret et al., 2016) dataset. See Appendix E for more details. The code for all the experiments presented in the paper will be available at github.com/naver/gdc.

4.3 Results

We present the evolution of our metrics through training epochs in Figure 3 (aggregated over tasks 1-6) and Figure 6 in the Appendix (aggregated over tasks 7-10). Results for each task are presented separately on Figures 7-10 in the Appendix.

Consistent with prior work (Khalifa et al., 2021; Korbak et al., 2022a), we observe that Reinforce is able to quickly achieve high levels of constraint satisfaction, but at the cost of large deviations from a, which translates into significantly decreased diversity of generated samples (in terms of Self-BLEU-5 and Distinct-1). The KL penalty term in Ziegler imposes an upper bound on deviation from a but the deviation is still significant enough to result in a drop in diversity. Moreover, we have observed Ziegler's objective to result in very unstable training.

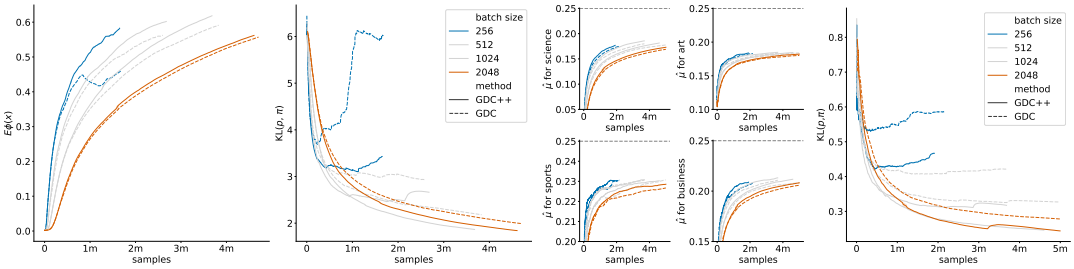
GDC and GDC++ are the only fine-tuning methods that address constraint satisfaction based on a clear formal objective, i.e. reducing the divergence from p. The approach translates into significantly smaller deviations from a and maintaining diversity within and across samples. The addition of a baseline indeed reduces the variance. We analyze that extensively in Appendix 4.5 while here focusing on the downstream effects of variance reduction. One is that π_θ is now able to compound staying closer to p and a *at the same time*, while achieving slightly better constraint satisfaction. We have also observed that baseline stabilizes training, leading to smoother curves.[9]

4.4 The effect of baseline across batch sizes

We expect that reducing gradient estimates variance can allow to train with lower batch sizes, performing gradient updates on estimates based on smaller batch sizes can increase the sample

[8]See Appendix D for a detailed description of how $D_{\mathrm{KL}}(p, \pi_\theta)$ is computed.

[9]The interested reader can compare the large fluctuations of the Ziegler objective to more stable training curves of GDC , and even more of GDC++ , in the disaggregated curves in Figures 7-10 of the Appendix.



(a) Task 1: a pointwise constraint

(b) Task 8: a set of distributional constraints; $\bar{\mu}_i = 0.25$

Figure 4: $\mathbb{E}_{\pi_\theta}\phi(x)$ or $\hat{\mu}$ per constraint (↑ better) and $D_{\mathrm{KL}}(p, \pi_\theta)$ (↓ better) as a function of the number of samples reported for task 1 (a) and task 8 (b). We report the number of samples (i.e. the number of epochs times the batch size) for a fair comparison of convergence speed. *GDC++ is consistently superior across all batch sizes in terms of convergence and constraint satisfaction*. The effect is more conspicuous with small batch sizes. Batch sizes 512 and 2014 are greyed out for clarity.

efficiency. To test this, we rerun tasks 1 (a pointwise constraint on the word "amazing") and 8 (distributional constraints on topics) with four batch sizes (256, 512, 1024, 2048). Figures 4a and 4b show the benefits of adding a baseline — higher constraint satisfaction, lower divergence from p, more stable training — and is especially evident with lower batch sizes. For instance, with batch size 256, GDC++ obtains a significantly higher constraint satisfaction rate and lower divergence from p.

Furthermore, stable training with smaller batch sizes translates into better sample efficiency. For instance, in task 1 (Figure 4a), GDC++ with batch size 256 needs 1M samples to achieve $\mathbb{E}_{x\sim\pi_\theta}\phi(x) = 0.5$ while GDC++ with batch size 2048 needs 4M. In contrast, GDC with batch size 256 does not achieve $\mathbb{E}_{x\sim\pi_\theta}\phi(x) = 0.5$ at all, confirming the importance of adding the baseline.

4.5 Empirical Evaluation of Variance Reduction

Next, we evaluate empirically the effect of the baseline for variance reduction. We select two tasks: task 1 (a pointwise constraint) and task 7 (distributional constraints) described in Section 4.2, each with 3 different seeds, while monitoring the following variance measures:

Gradient Variance The gradient estimate is defined as: $G_\theta(x) \doteq A(x)\nabla_\theta \log \pi_\theta(x)$, where $G_\theta(x) \in \mathbb{R}^{|\theta|}$ is an unbiased estimate of the gradient of the forward KL loss $\nabla_\theta D_{\mathrm{KL}}(p, \pi_\theta)$ with respect to the parameters θ. We then have, with $\mu(G_\theta) \doteq \mathbb{E}_{x\sim q} G_\theta(x)$:

$$\mathrm{Var}(G_\theta) \doteq \mathbb{E}_{x\sim q} \|G_\theta(x) - \mu(G_\theta)\|_2^2 \tag{17}$$

$$= \mathbb{E}_{x\sim q}||G_\theta(x)||_2^2 - ||\mu(G_\theta)||_2^2. \tag{18}$$

Variance of the advantage is defined by:

$$\mathrm{Var}\,(\mathrm{A}) \doteq \mathbb{E}_{x\sim q} \left\|\mathrm{A}(x) - \mu^{\mathrm{A}}\right\|_2^2 \tag{19}$$

where, $\mu^{\mathrm{A}} \equiv \mathbb{E}_{x\sim q}\ \mathrm{A}(x)$ is the mean of the advantage, which we showed above to be null after the addition of the baseline.

Expected absolute value of the advantage This metric is defined as:

$$\mu^{|\mathrm{A}|} \doteq \mathbb{E}_{x\sim q}\ |\mathrm{A}(x)|. \tag{20}$$

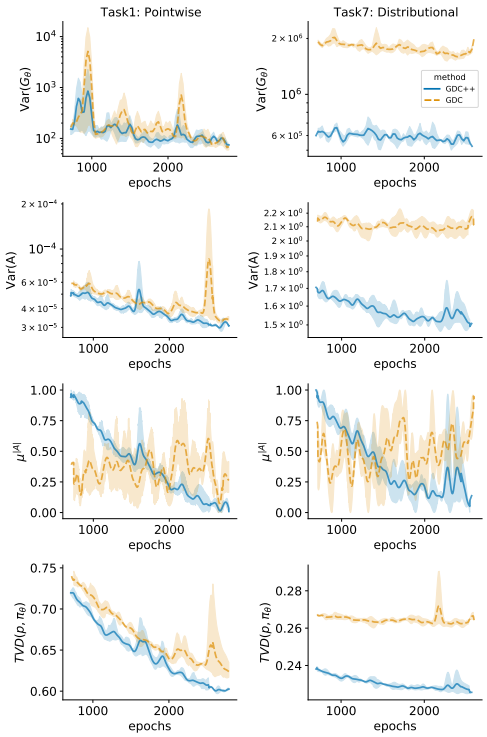


Figure 5: Comparison between GDC and GDC++ using a set of Variance diagnosis metrics on pointwise and distributional constraints experiments.

It directly provides a standard measure of distributional discrepancy between p and π_θ, in terms of TVD (Total Variation Distance). We have:

$$\mathbb{E}_{x\sim q}\left|\frac{p(x)}{q(x)}-\frac{\pi_\theta(x)}{q(x)}\right| = 2\,\mathrm{TVD}(p,\pi_\theta). \tag{21}$$

Results Figure 5 shows that GDC++ obtains lower variance in the gradient estimates $\mathrm{Var}(G_\theta)$ and the variance of the advantage $\mathrm{Var}\,(\mathrm{A})$ in both pointwise and distributional experiments compared to its non-baseline counterpart GDC. We further observe a decreasing trend in the mean absolute value of the advantage $\mu^{|\mathrm{A}|}$ which is correlated with a decreasing trend in the TVD distance between the trained policy π_θ and the optimal distribution p. Overall, these results shows that adding a baseline to DPG reduces the variance during training and yields better convergence towards the optimal distribution p.

5 Related work

The idea of posing control problems as distribution matching has resurfaced numerous times in the RL literature (Kappen et al., 2012; Friston et al., 2010; Levine, 2018; Hafner et al., 2020; Buckley et al., 2017). KL-control can be seen as a generalisation of maximum entropy RL (MaxEnt RL) (Haarnoja et al., 2017, 2018) to informed priors. If in (2) we chose $a(x)$ to be a uniform distribution (assuming right now finiteness of X) instead of a pretrained LM distribution, then the KL penalty $D_{\mathrm{KL}}(\pi_\theta, a)$ would reduce to an entropy bonus. Both KL-control and MaxEnt RL can be derived from a general framework of control-as-inference (Levine, 2018) which poses control as minimising KL from a certain target distribution. However, most practical algorithms in the MaxEnt RL family minimise KL from a target policy which changes throughout training; in contrast, DPG's target distribution p and KL-control implicit target distribution p_z are defined at trajectory level and fixed throughout training.

Perhaps the closest method to KL-control and DPG in the larger family of inference-based RL (Furuta et al., 2021) is AWR (Peng et al., 2019) which minimises the *forward* KL from an off-policy target distribution. Yet another approach with apparent similarity to KL-control and DPG is state marginal matching (SMM) (Hazan et al., 2018; Lee et al., 2019). SMM poses exploration as learning a policy that induces a state marginal distribution that matches a target state distribution. While SMM's target distribution is fixed, it is defined for individual states, while in the controllable language generation tasks we consider, the target distribution is defined over a complete trajectory considered as a unit. See Appendix B for an extended discussion of related work.

6 Conclusion

Fine-tuning large language models has become an active area of research, due to its importance in adapting large language models to satisfy task-level preferences, or in combating their social risks such as "distributional" stereotyping (Weidinger et al., 2021; Welbl et al., 2021) [10]. In this paper, we analyzed in depth the nuanced relation between two popular fine-tuning paradigms: RM and DM. We demonstrated that KL-control can be seen as a form of DM and showed that while DPG and PG have different goals, some similarities (similar forms of gradient estimates despite different objectives) can be exploited. We used these insights to inform an extension of DPG, consisting in adding a baseline to reduce the variance of gradient estimates.

The connections we established suggest that despite fundamental differences between DPG and RL, some of the theoretical results and algorithmic techniques from RL can be adapted to a DM framework without losing their formal guarantees. In this paper, we focus on variance reduction using baselines, but the space of possible enhancements is vast. Promising candidates include further reducing the variance using a learned value function (Konda & Tsitsiklis, 2000) and preventing detrimentally large policy updates by maintaining a trust region in the policy space – akin to techniques such as TRPO (Schulman et al., 2015) and PPO (Schulman et al., 2017b). Another future direction could consist in analyzing the relation between explicit EBMs in DPG and implicit EBMs arising in KL-control and characterizing the space of EBMs that could be reached through KL-control.

[10] See Appendix A for a discussion of broader impacts of large language models and controllable language generation.

References

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally Normalized Transition-Based Neural Networks. 2016. doi: 10.18653/v1/P16-1231.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. An Actor-Critic Algorithm for Sequence Prediction. (2015):1–17, 2016. URL http://arxiv.org/abs/1607.07086.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. An actor-critic algorithm for sequence prediction. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJDaqqveg.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

A. Bakhtin, Y. Deng, S. Gross, Myle Ott, Marc'Aurelio Ranzato, and Arthur Szlam. Energy-based models for text. *ArXiv*, abs/2004.10188, 2020.

David Belanger and Andrew McCallum. Structured prediction energy networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 983–992. JMLR.org, 2016. URL http://dl.acm.org/citation.cfm?id=3045390.3045495.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://www.aclweb.org/anthology/2020.acl-main.485.

Christopher L. Buckley, Chang Sub Kim, Simon McGregor, and Anil K. Seth. The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81:55–79, 2017. ISSN 0022-2496. doi: https://doi.org/10.1016/j.jmp.2017.09.004. URL https://www.sciencedirect.com/science/article/pii/S0022249617300962.

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. Language gans falling short. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJgza6VtPB.

Leshem Choshen, Lior Fox, Zohar Aizenbud, and Omri Abend. On the weaknesses of reinforcement learning for neural machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=H1eCw3EKvH.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Pre-training transformers as energy-based cloze models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 285–294. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.20. URL https://doi.org/10.18653/v1/2020.emnlp-main.20.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=H1edEyBKDS.

Peter Dayan. Reinforcement comparison. In *Proceedings of the 1990 Connectionist Models Summer School*, pp. 45–51. Morgan Kaufmann, San Mateo, CA, 1990.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. Residual energy-based models for text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=B1l4SgHKDH.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Karl J Friston, Jean Daunizeau, James Kilner, and Stefan J Kiebel. Action and behavior: a free-energy formulation. *Biological cybernetics*, 102(3):227–260, 2010.

Hiroki Furuta, Tadashi Kozuno, Tatsuya Matsushima, Yutaka Matsuo, and Shixiang Shane Gu. Co-adaptation of algorithmic and implementational innovations in inference-based deep reinforcement learning, 2021. URL https://arxiv.org/abs/2103.17258.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *J. Mach. Learn. Res.*, 5:1471–1530, December 2004. ISSN 1532-4435.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 297–304, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL http://proceedings.mlr.press/v9/gutmann10a.html.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1352–1361. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/haarnoja17a.html.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018. URL https://arxiv.org/abs/1801.01290.

Danijar Hafner, Pedro A. Ortega, Jimmy Ba, Thomas Parr, Karl Friston, and Nicolas Heess. Action and perception as divergence minimization, 2020.

Elad Hazan, Sham M. Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration, 2018. URL https://arxiv.org/abs/1812.02690.

Tianxing He, Bryan McCann, Caiming Xiong, and Ehsan Hosseini-Asl. Joint energy-based model training for better calibrated natural language understanding models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1754–1761, Online, April 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.eacl-main.151.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800, 2002. doi: 10.1162/089976602760128018. URL https://doi.org/10.1162/089976602760128018.

Natasha Jaques, Shixiang Gu, Dzmitry Bahdanau, José Miguel Hernández-Lobato, Richard E. Turner, and Douglas Eck. Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1645–1654. PMLR, 2017a. URL http://proceedings.mlr.press/v70/jaques17a.html.

Natasha Jaques, Shixiang Gu, Dzmitry Bahdanau, Jose Miguel Hernandez Lobato, Richard E. Turner, and Doug Eck. Tuning recurrent neural networks with reinforcement learning. 2017b. URL https://openreview.net/pdf?id=Syyv2e-Kx.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Àgata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind W. Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *CoRR*, abs/1907.00456, 2019. URL http://arxiv.org/abs/1907.00456.

Hilbert J Kappen, Vicenç Gómez, and Manfred Opper. Optimal control as a graphical model inference problem. *Machine learning*, 87(2):159–182, 2012.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. A distributional approach to controlled text generation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jWkw45-9AbL.

Samuel Kiegeland and Julia Kreutzer. Revisiting the weaknesses of reinforcement learning for neural machine translation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 1673–1681. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.133. URL https://doi.org/10.18653/v1/2021.naacl-main.133.

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

A.S. Klyubin, D. Polani, and C.L. Nehaniv. Empowerment: a universal agent-centric measure of control. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pp. 128–135 Vol.1, 2005. doi: 10.1109/CEC.2005.1554676.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000. URL https://proceedings.neurips.cc/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Tomasz Korbak, Hady Elsahar, Marc Dymetman, and Germán Kruszewski. Energy-based models for code generation under compilability constraints. *CoRR*, abs/2106.04985, 2021. URL https://arxiv.org/abs/2106.04985.

Tomasz Korbak, Hady Elsahar, German Kruszewski, and Marc Dymetman. Controlling conditional language models without catastrophic forgetting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11499–11528. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/korbak22a.html.

Tomasz Korbak, Ethan Perez, and Christopher L Buckley. RL with KL penalties is better viewed as Bayesian inference, 2022b. URL https://arxiv.org/abs/2205.11275.

Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In Jian Su, Xavier Carreras, and Kevin Duh (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 1203–1213. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/d16-1128. URL https://doi.org/10.18653/v1/d16-1128.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A Tutorial on Energy-Based Learning. In *Predicting Structured Data*. MIT Press, 2006.

Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. 2019.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review, 2018.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, San Diego, California, June 2016a. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL https://www.aclweb.org/anthology/N16-1014.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In Jian Su, Xavier Carreras, and Kevin Duh (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 1192–1202. The Association for Computational Linguistics, 2016b. doi: 10.18653/v1/d16-1127. URL https://doi.org/10.18653/v1/d16-1127.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models, 2021.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In Jian Su, Xavier Carreras, and Kevin Duh (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 2122–2132. The Association for Computational Linguistics, 2016a. doi: 10.18653/v1/d16-1230. URL https://doi.org/10.18653/v1/d16-1230.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Optimization of image description metrics using policy gradient methods. *CoRR*, abs/1612.00370, 2016b. URL http://arxiv.org/abs/1612.00370.

Beren Millidge, Alexander Tschantz, Anil Seth, and Christopher Buckley. Understanding the origin of information-seeking exploration in probabilistic objectives for control, 2021.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1791–1799, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/mnih14.html.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013. URL http://arxiv.org/abs/1312.5602.

Subhajit Naskar, Pedram Rooshenas, Simeng Sun, Mohit Iyyer, and A. McCallum. Energy-based reranking: Improving neural machine translation using energy-based models. *ArXiv*, abs/2009.13267, 2020.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Art B. Owen. Importance Sampling. In *Monte Carlo theory, methods and examples*, chapter 9. 2013. URL https://statweb.stanford.edu/~owen/mc/Ch-var-is.pdf.

Tetiana Parshakova, Jean-Marc Andreoli, and Marc Dymetman. Global Autoregressive Models for Data-Efficient Sequence Learning. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 900–909, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/K19-1084. URL https://www.aclweb.org/anthology/K19-1084.

Tetiana Parshakova, Jean-Marc Andreoli, and Marc Dymetman. Distributional Reinforcement Learning For Energy-Based Sequential Models. *CoRR*, 2019b. URL https://arxiv.org/abs/1912.08517.

Ramakanth Pasunuru and Mohit Bansal. Reinforced video captioning with entailment rewards. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 979–985. Association for Computational Linguistics, 2017. doi: 10.18653/v1/d17-1103. URL https://doi.org/10.18653/v1/d17-1103.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=HkAClQgA-.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning, 2019. URL https://arxiv.org/abs/1910.00177.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models, 2022. URL https://arxiv.org/abs/2202.03286.

Jan Peters and Stefan Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21(4):682–697, 2008. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2008.02.003. URL https://www.sciencedirect.com/science/article/pii/S0893608008000701. Robotics and Neuroscience.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

Marc'Aurelio Ranzato, Y-Lan Boureau, Sumit Chopra, and Yann LeCun. A unified energy-based framework for unsupervised learning. In Marina Meila and Xiaotong Shen (eds.), *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, AISTATS 2007, San Juan, Puerto Rico, March 21-24, 2007*, volume 2 of *JMLR Proceedings*, pp. 371–379. JMLR.org, 2007. URL http://proceedings.mlr.press/v2/ranzato07a.html.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.06732.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017a. URL http://arxiv.org/abs/1707.06347.

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

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 3405–3410. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1339. URL https://doi.org/10.18653/v1/D19-1339.

David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103535. URL https://www.sciencedirect.com/science/article/pii/S0004370221000862.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize from human feedback. *CoRR*, abs/2009.01325, 2020. URL https://arxiv.org/abs/2009.01325.

Richard S. Sutton. *Temporal credit assignment in reinforcement learning*. PhD thesis, University of Massachusetts, 1984.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pp. 1057–1063, Cambridge, MA, USA, 1999. MIT Press.

Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J. Martin, Animesh Mehta, Brent Harrison, and Mark O. Riedl. Controllable neural story plot generation via reward shaping. In Sarit Kraus (ed.), *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pp. 5982–5988. ijcai.org, 2019. doi: 10.24963/ijcai.2019/829. URL https://doi.org/10.24963/ijcai.2019/829.

Emanuel Todorov. Linearly-solvable markov decision problems. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007. URL https://proceedings.neurips.cc/paper/2006/file/d806ca13ca3449af72a1ea5aedbed26a-Paper.pdf.

Lifu Tu, Richard Yuanzhe Pang, Sam Wiseman, and Kevin Gimpel. Engine: Energy-based inference networks for non-autoregressive machine translation. *ArXiv*, abs/2005.00850, 2020.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 4566–4575. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7299087. URL https://doi.org/10.1109/CVPR.2015.7299087.

Lex Weaver and Nigel Tao. The optimal reward baseline for gradient-based reinforcement learning. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pp. 538–545, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359, 2021. URL https://arxiv.org/abs/2112.04359.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. Challenges in detoxifying language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pp. 2447–2469. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.210. URL https://doi.org/10.18653/v1/2021.findings-emnlp.210.

Ronald J. Williams. Reinforcement-learning connectionist systems. Technical report, Northeastern University, 1987. Technical Report NU-CCS-87-3.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8:229–256, 1992a. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Machine Learning*, pp. 229–256, 1992b.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019. URL http://arxiv.org/abs/1910.03771.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL http://arxiv.org/abs/1609.08144.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (eds.), *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pp. 1097–1100. ACM, 2018. doi: 10.1145/3209978.3210080. URL https://doi.org/10.1145/3209978.3210080.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019. URL http://arxiv.org/abs/1909.08593.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix A.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] In Appendix C we present proofs of all mathematical facts referred to in the paper.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code is included as supplementary material available to the reviewers and area chairs and will be made publicly available alongside the camera ready version of the paper.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In Appendix E we provide the hyperparameters used throughout our experiments and report our hardware configuration. In Appendix D, we describe in detail how $D_{\mathrm{KL}}(p, \pi_\theta)$ and $\mathrm{TVD}(p, \pi_\theta)$ were estimated and provide an extended pseudocode for our training loop in Algorithm 2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] However, we found the variance across random seeds to be negligible and not comparing across random seeds is a standard practice when working with large language models where the cost of a single run is significant.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In Appendix E we report our hardware configuration.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] In Appendix E.
   (b) Did you mention the license of the assets? [Yes] In Appendix E.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Broader impacts

The focus area of this paper — fine-tuning large language models — is aligned with an important line of work on addressing the problem of social bias in large language models (Sheng et al., 2019; Liang et al., 2021). As the training data for large language models consists mainly of crawled user-generated content, a number of factors (from crawling methodology to Internet participation inequalities and moderation practices) leads to an over-representation of certain viewpoints and voices exceeding their prevalence in the general population. This poses a risk of amplifying biases and harms through a language model perpetuating these voices (Bender et al., 2021; Blodgett et al., 2020; Sheng et al., 2019; Weidinger et al., 2021; Welbl et al., 2021). Numerous problems related to addressing data bias in language generation (e.g. controlling for gender distribution in generated texts) can be naturally posed as generative distributional control (GDC), the framework we focus our experiments on. The *distributional* character of these data bias problems lies in the fact that desirable properties of generated texts are defined for a collection of samples, not only for individual samples. Our theoretical analyses of reward maximization and distribution matching approaches as well as our algorithmic improvements to the GDC framework — termed GDC++ — are therefore also a contribution to the problem of bias in language models. However, we need to be aware that GDC++ , KL-control as well as controllable language generation techniques in general, can also be diverted to malicious uses such as spreading misinformation or generating harmful content.

B Extended Related Work

Reinforcement learning for language generation Most previous attempts at steering language models to conform to global constraints defined over entire sequences have employed reinforcement learning. This includes using Reinforce (Williams, 1992a) for machine translation Ranzato et al. (2016), actor critic (Konda & Tsitsiklis, 2000) for abstractive summarization (Paulus et al., 2018), caption generation (Liu et al., 2016b), dialogue (Li et al., 2016b), and video captioning (Pasunuru & Bansal, 2017). Some approaches (for instance, in machine translation and summarization (Ranzato et al., 2016; Bahdanau et al., 2017)) directly optimize performance metrics such as BLEU and ROUGE at training time. Others use heuristic rewards (for instance Li et al. (2016b) for dialogue generation and Tambwekar et al. (2019) for story generation) in order to obtain certain *a priori* desirable features of generated sequences that then incentivize good performance on target metrics. Catastrophic forgetting is a frequent problem of these fine-tuning approaches: reward maximization happens at the expense of large deviations from the original model. This problem is sometimes addressed by imposing a penalty term to the rewards, such as the KL divergence between the trained policy and the auto-regressive model. This approach, termed "conservative fine-tuning", was applied to generating melodies with music theory rewards and organic molecules with synthesizability rewards by Jaques et al. (2017a) as well fine-tuning language models for controllable language generation by Ziegler et al. (2019). This solution often has hard time balancing between the reward term and the KL penalty term, leading to instability in training (Khalifa et al., 2021; Korbak et al., 2022a). Unlike this approach, KL-DPG determines an optimal distribution that satisfies both requirements.

RM and DM objectives in control problems While RM is the dominant approach to tackling control problems (Sutton & Barto, 2018) and is sometimes argued to be sufficient for any intelligent behavior (Silver et al., 2021), prior work explored the benefits of alternative objectives formulated as DM: minimizing divergence from some target distribution p. Prominent examples of (families of) DM objectives include control state marginal matching (Lee et al., 2019) active inference (Friston et al., 2010; Buckley et al., 2017) and control-as-inference (Kappen et al., 2012; Todorov, 2007; Levine, 2018). Hafner et al. (2020) propose a *reverse* KL from a joint distribution over observations and latent variables as a universal objective for action and perception that — depending on a choice of the target p — gives rise to many familiar objectives, including empowerment (Klyubin et al., 2005), maximum entropy RL (Haarnoja et al., 2017) or KL-control (Todorov, 2007). In a similar vein, Millidge et al. (2021) compare RM and DM objectives (or, evidence and divergence objectives, according to their terminology) in the context of exploration. They conclude that information-seeking exploration arises naturally in DM but *not* in RM. This is because, when the target distribution p involves latent variables, a DM objective decomposes into an information gain term that pushes the agent to seek observations that are most informative of latent variables. In contrast, RM objectives entail *minimizing* information gain between latent variables and observations. Finally, (Korbak et al.,

2022b) defend an interpretation of KL-control for controlling language models as Bayesian inference: updating a prior a to conform to evidence provided by a reward function R.

Maximum entropy RL Maximum entropy RL (MaxEnt RL)'s objective is maximising expected reward minus policy entropy. KL-control can be seen as generalisation of maximum-entropy RL (Haarnoja et al., 2017, 2018) to informed priors. If in (2) we chose $a(x)$ to be a uniform distribution (an uninformed prior) instead of a pretrained LM distribution, then the KL penalty $D_{\mathrm{KL}}(\pi_\theta, a)$ would reduce to an entropy bonus and KL-control's objective would reduce to a standard Maximum entropy RL objective. Both KL-control and Maximum entropy RL can be derived from a general framework of control-as-inference (Levine, 2018) which poses control as minimising KL from a certain target distribution. However, while KL-control (Ziegler et al., 2019) and DPG directly minimise a single KL from a target distribution over whole sequences (trajectories), most practical algorithms in the maximum entropy family RL approximate it by related but importantly different objectives.

The three biggest differences between MaxEnt RL on the one hand and DPG and KL-control (Ziegler et al., 2019) on the other hand are as follows:

1. KL-control implicit target distribution p_z and DPG's target distribution p are over whole sequences (trajectories) while in most MaxEnt RL algorithms the target distribution over actions conditioned on a state: $\pi^*(a|s)$. For instance in both SQL (Haarnoja et al., 2017) and SAC (Haarnoja et al., 2018) the target distribution is defined as $\pi^*(a|s) = \exp(Q_\theta(s,a))/Z_\theta(s)$, where Q is a state-action value function and Z is a partition function of for a given state, both dependent on policy parameters θ.
2. KL-control's implicit target distribution and DPG's target distribution are predefined (i.e. held constant throughout training). In MaxEnt RL it typically undergoes updates. Again, in both SQL (Haarnoja et al., 2017) and SAC (Haarnoja et al., 2018) they depend on a Q function which is continuously updated on new trajectories.
3. KL-control's implicit target distribution p_z and DPG's target distribution p involve an informed prior $a(x)$: a pretrained language model. In most MaxEnt RL algorithms, the prior is assumed to be a uniform distribution.

Because MaxEnt RL algorithms do not approximate a constant, predefined target distribution, they cannot be framed as minimising a single KL objective. Instead, they typically implement (soft) policy iteration (Sutton & Barto, 2018): they alternate between defining a new target distribution (policy evaluation) and minimising KL from that current target distribution (policy improvement). In other words, minimising KL is a subroutine of policy iteration, not an objective in itself.

Perhaps the closest method to KL-control and DPG in the larger family of inference-based RL (Furuta et al., 2021) is AWR (Peng et al., 2019), which minimises the *forward* KL from a target distribution $\frac{1}{Z}\mu(a|s)\exp(A(s,a))$, where μ is a behavioural policy implicitly defined by the trajectory buffer and A is the advantage. Here, the prior is informative and given by the policy from a previous iteration k. However, the target distribution is not constant: it is updated on each iteration.

State marginal matching State marginal matching (Hazan et al., 2018; Lee et al., 2019) is an approach to exploration in RL. It poses exploration as learning a policy π that induces a state marginal distribution $\rho_\pi(s) = \mathbb{E}\sum_{t=1}^{T} \mathbb{1}(s_t = t)$ that matches a given target state distribution p^*. While this approach differs in motivation from DPG and KL-control (it solves the problem of exploration in the space of policies, not constraint satisfaction), it optimises a similar divergence objective: $D_{\mathrm{KL}}(\pi, p^*)$. Unlike in maximum-entropy RL, the target p^* is fixed. However, p^* is a distribution over states, not trajectories (as in the case of p in DPG and p_z in KL-control). There is no obvious notion of state in the controllable language generation tasks we consider other than treating the whole sequence as a state.

Baselines in Reinforcement Learning In the context of reinforcement learning, baselines were introduced by Sutton (1984). Williams (1987, 1992a) has shown them to reduce variance in a number of use cases and also proved that they do not introduce bias. Dayan (1990) was the first to observe and confirm experimentally that the optimal constant baseline is not equal to expected reward in a simple two-arm bandit setting. This result was generalized to POMDPs (Partially Observable Markov Decision Processes) by Weaver & Tao (2001, section 3.1.3, p. 540) and variable baselines by

Greensmith et al. (2004, theorem 13, p. 1489) who also proved bounds on the variance of gradient estimates. The optimal baseline, however, is rarely used in practice (Sutton & Barto (2018); for an exception, see (Peters & Schaal, 2008)). Outside RL, baselines were also used in the context of learning inference networks for amortized variational inference by Mnih & Gregor (2014) and found to yield similar variance reduction.

Energy-based models for language Energy-based models (EBMs) (Hinton, 2002; LeCun et al., 2006; Ranzato et al., 2007) are a family of models in which learning and inference are done by associating an unnormalized probability with each configuration of observed and latent variables. Early examples of EBMs applied to natural language processing include sequence labeling problems (e.g. tagging) exploiting global properties of a sequence (Andor et al., 2016; Belanger & McCallum, 2016). The recent surge of interest in EBMs has not left natural language processing unaffected (see Bakhtin et al. (2020) for a survey). Tu et al. (2020) proposed an energy-based inference networks for non-autoregressive machine translation while Naskar et al. (2020) use an EBM for reranking candidate translations according to their predicted BLEU scores. Parshakova et al. (2019a) and Deng et al. (2020) augment an autoregressive language models with an additional global factor to obtain a lower perplexity on the training data. Clark et al. (2020) poses non-autoregressive language modeling as training an energy-based cloze task scorer using noise-contrastive estimation (Gutmann & Hyvärinen, 2010). He et al. (2021) obtain better calibration on natural language inference tasks by augmenting and training the classifier jointly with an energy-based model modeling the marginal distribution over samples, again using noise-contrastive estimation. In consequence, the classifier tends to assign more conservative (high-entropy) predictions to high-energy (less likely, possibly out of distribution) samples.

C Additional proofs

C.1 Optimal baselines in RL

Despite its widespread use, the baseline as mean of reward

$$B^{\text{RL}} = \mathbb{E}_{x\sim\pi_\theta(x)} R(x) \tag{22}$$

is not the optimal constant baseline for reward maximization objectives in RL. The optimal constant baseline, i.e. one yielding the minimal variance of the gradient, is given by:

$$B^* = \frac{\mathbb{E}_{x\sim\pi_\theta}[R(x)\left(\nabla_\theta \log \pi_\theta(x)\right)^2]}{\mathbb{E}_{x\sim\pi_\theta}[\left(\nabla_\theta \log \pi_\theta(x)\right)^2]}. \tag{23}$$

In order to maintain accessibility, in this section, we provide a self-contained derivation of this optimal form of baselines (23) and and connect it to the commonly used form (22).[11]

First, recall that $R(x)$ is a reward associated with an input x. B is a baseline value subtracted from the reward that does not introduce bias in gradient estimation. Now let's denote the gradient wrt an individual sample x as $G_\theta(x)$ where

$$G_\theta(x) = [R(x) - B]\nabla_\theta \log \pi_\theta(x), \tag{24}$$

and the estimate of the gradient as

$$G(\theta) = \mathbb{E}_{x\sim\pi_\theta} G_\theta(x). \tag{25}$$

Using the general identity $\mathbf{var}(z) = \mathbb{E}[z^2] - [\mathbb{E}z]^2$, the variance of the gradient takes the form:

$$\text{Var}(G_\theta) = \mathbb{E}_{x\sim\pi_\theta}[G_\theta(x)^2] - G(\theta)^2 \tag{26}$$

Now let's take the gradient of this variance with respect to B and solve to find the baseline form with minimal variance:

$$\frac{d\text{Var}(G_\theta)}{dB} = \frac{d}{dB}\mathbb{E}_{x\sim\pi_\theta}[(G_\theta(x))^2] - \frac{d}{dB}(\mathbb{E}_{x\sim\pi_\theta}[G_\theta(x)])^2. \tag{27}$$

[11] The formula for the optimal baseline in (23) was originally proved by Weaver & Tao (2001) but here we provide a simpler proof sketched by Sergey Levine in his slides: `http://rail.eecs.berkeley.edu/deeprlcourse-fa17/f17docs/lecture_4_policy_gradient.pdf`

The second term of the right hand side of (27) is equal to zero, since B does not introduce bias into $G(\theta)$:

$$\frac{d}{dB}\left(\mathbb{E}_{x\sim\pi_\theta}[G_\theta(x)]\right)^2 = \frac{d}{dB}\left(\mathbb{E}_{x\sim\pi_\theta}\left[(R(x)-B)\nabla_\theta \log \pi_\theta(x)\right]\right)^2$$
$$= \frac{d}{dB}\left(\mathbb{E}_{x\sim\pi_\theta}\left[R(x)\nabla \log \pi_\theta(x)\right]\right)^2 = 0.$$

Plugging this back into (27), we obtain:

$$\begin{aligned}\frac{d\mathrm{Var}(G_\theta)}{dB} &= \frac{d}{dB}\mathbb{E}_{x\sim\pi_\theta}[(G_\theta(x))^2] \\ &= \mathbb{E}_{x\sim\pi_\theta}\left[\frac{d}{dB}\left[\left(R(x)^2+B^2-2R(x)B\right)\left(\nabla_\theta \log \pi_\theta(x)\right)^2\right]\right] \\ &= \mathbb{E}_{x\sim\pi_\theta}(2B-2R(x))(\nabla_\theta \log \pi_\theta(x))^2 \\ &= 2B\,\mathbb{E}_{x\sim\pi_\theta}(\nabla_\theta \log \pi_\theta(x))^2 - 2\,\mathbb{E}_{x\sim\pi_\theta}R(x)\left(\nabla_\theta \log \pi_\theta(x)\right)^2.\end{aligned}$$

Then, solving $\frac{d\mathrm{Var}(G_\theta)}{dB}=0$ for B, we obtain the optimal form of the baseline B^* as required:

$$B^* = \frac{\mathbb{E}_{x\sim\pi_\theta}[R(x)\left(\nabla_\theta \log \pi_\theta(x)\right)^2]}{\mathbb{E}_{x\sim\pi_\theta}[\left(\nabla_\theta \log \pi_\theta(x)\right)^2]}. \tag{28}$$

This can be interpreted as average reward (as in B^{RL}) but weighted by gradient magnitudes $(\nabla_\theta \log \pi_\theta(x))^2$. Moreover, $B^* = B^{\mathrm{RL}}$ is recovered **under the condition that** the reward $R(x)$ is uncorrelated (*a fortiori* independent) from $(\nabla_\theta \log \pi_\theta(x))^2$. If that were the case, we would have:

$$B^* = \frac{\mathbb{E}_{x\sim\pi_\theta}[R(x)\left(\nabla_\theta \log \pi_\theta(x)\right)^2]}{\mathbb{E}_{x\sim\pi_\theta}[\left(\nabla_\theta \log \pi_\theta(x)\right)^2]} \tag{29}$$

$$= \frac{\mathbb{E}_{x\sim\pi_\theta}[R(x)]\;\mathbb{E}_{x\sim\pi_\theta}[\left(\nabla_\theta \log \pi_\theta(x)\right)^2]}{\mathbb{E}_{x\sim\pi_\theta}[\left(\nabla_\theta \log \pi_\theta(x)\right)^2]} \tag{30}$$

$$= \mathbb{E}_{x\sim\pi_\theta}[R(x)] = B^{\mathrm{RL}}. \tag{31}$$

C.2 unbiasedness of PG baseline

Baselines are a standard variance reduction technique in the context of Policy Gradients (Sutton & Barto, 2018). The idea is to subtract from the reward $R(x)$ a value B that does not introduce bias to the gradients but may change variance. Equation (1) then takes the following form:

$$\nabla_\theta \mathbb{E}_{\pi_\theta} R(x) = \mathbb{E}_{\pi_\theta}(R(x)-B)\,\nabla_\theta \log \pi_\theta(x). \tag{32}$$

To see that B does not introduce bias, we can rewrite (11) as:

$$\mathbb{E}_{x\sim\pi_\theta} R(x)\nabla_\theta \log \pi_\theta(x) - B\,\mathbb{E}_{\pi_\theta}\nabla_\theta \log \pi_\theta(x) \tag{33}$$

and note that the second term is null because $\sum_x \pi_\theta(x)\nabla_\theta \log \pi_\theta(x) = \nabla_\theta \sum_x \pi_\theta(x) = 0$.

C.3 Unbiasedness of DPG Baseline

Recall that the gradient estimate for DPG (Parshakova et al., 2019a) has the following form:

$$\mathbb{E}_{x\sim\pi_\theta}\frac{P(x)}{\pi_\theta(x)}\nabla_\theta \log \pi_\theta(x) \tag{34}$$

After subtracting a baseline $B = Z$, it becomes

$$\mathbb{E}_{x\sim\pi_\theta}\Big[\frac{P(x)}{\pi_\theta(x)} - Z\Big]\nabla_\theta \log \pi_\theta(x) = \mathbb{E}_{x\sim\pi_\theta}\frac{P(x)}{\pi_\theta(x)}\nabla_\theta \log \pi_\theta(x) - Z\Big[\mathbb{E}_{x\sim\pi_\theta}\nabla_\theta \log \pi_\theta(x)\Big] \tag{35}$$

$$= \mathbb{E}_{x\sim\pi_\theta}\frac{P(x)}{\pi_\theta(x)}\nabla_\theta \log \pi_\theta(x) - Z\Big[\sum_x \nabla_\theta \pi_\theta(x)\Big] \tag{36}$$

Here, the second term does not introduce bias because $Z\Big[\sum_x \nabla_\theta \pi_\theta(x)\Big] = 0$, leaving us with the same exact form of gradient as in the DPG algorithm.

C.4 Unbiasedness of DPG^{off} baseline

Offline DPG, the off policy variant of DPG proposed in Parshakova et al. (2019b); Khalifa et al. (2021) has the following gradient estimate:

$$\mathbb{E}_{x\sim q}\frac{P(x)}{q(x)}\nabla_\theta \log \pi_\theta(x) \tag{37}$$

Where q is a proposal distribution (another auto-regressive model) used to detach the training of π_θ from the sampling process and allow more stable training.

Recall that the Baseline of DPG^{off} is of the form:

$$B^{\text{off}}(x) = Z\frac{\pi_\theta(x)}{q(x)}, \tag{38}$$

The $\frac{\pi_\theta(x)}{q(x)}$ term is an importance weight correcting for the bias introduced by sampling from q.

Unbiasedness To show that subtracting a baseline $B^{\text{off}}(x) = Z\frac{\pi_\theta(x)}{q(x)}$ doesn't introduce bias, let's rewrite the gradient estimate with added baseline as a sum of two terms:

$$\mathbb{E}_{x\sim q}\Big[\frac{P(x)}{q(x)} - Z\frac{\pi_\theta(x)}{q(x)}\Big]\nabla_\theta \log \pi_\theta(x) = \Big[\mathbb{E}_{x\sim q}\frac{P(x)}{q(x)}\nabla_\theta \log \pi_\theta\Big] - \Big[\mathbb{E}_{x\sim q}Z\frac{\pi_\theta(x)}{q(x)}\nabla_\theta \log \pi_\theta\Big] \tag{39}$$

$$= \Big[\mathbb{E}_{x\sim q}\frac{P(x)}{q(x)}\nabla_\theta \log \pi_\theta\Big] - Z\Big[\sum_x \nabla_\theta \pi_\theta(x)\Big] \tag{40}$$

Here again the second term does not introduce bias because $Z\Big[\sum_x \nabla_\theta \pi_\theta(x)\Big] = 0$.

Null Advantage on Average In the case of sampling with π_θ in the online DPG choosing $B = Z$ had the benefit that the advantage $R_\theta(x) - B$ was centered around 0, namely: $\mathbb{E}_{x\sim\pi_\theta}[R_\theta(x) - Z] = 0$.

With the $B^{\text{off}}(x)$ baseline for the DPG^{off} this important property is also maintained. The advantage now takes the form $\frac{P(x)}{q(x)} - Z\frac{\pi_\theta(x)}{q(x)}$ and then:

$$\mathbb{E}_{x\sim q}\Big[\frac{P(x)}{q(x)} - Z\frac{\pi_\theta(x)}{q(x)}\Big] = \sum_x P(x) - Z\pi_\theta(x) \tag{41}$$

$$= Z - Z\sum_x \pi_\theta(x) = 0. \tag{42}$$

To visualize things better, we elaborate the difference in forms of rewards, baseline and gradients before and after addition of the baseline between DPG (on policy) and DPG^{off} (off policy) in Table 2.

	DPG	**DPG^{off}**
Reward	$\frac{P(x)}{\pi_\theta(x)}$	$\frac{P(x)}{q(x)}$
∇_θ	$\mathbb{E}_{x\sim\pi_\theta}\frac{P(x)}{\pi_\theta(x)}\nabla_\theta \log \pi_\theta(x)$	$\mathbb{E}_{x\sim q}\frac{P(x)}{q(x)}\nabla_\theta \log \pi_\theta(x)$
Baseline	Z	$Z\frac{\pi_\theta(x)}{q(x)}$
Advantage	$\frac{P(x)}{\pi_\theta(x)} - Z$	$\frac{P(x)}{q(x)} - Z\frac{\pi_\theta(x)}{q(x)}$
∇_θ **with baseline**	$\mathbb{E}_{x\sim\pi_\theta}\left[\frac{P(x)}{\pi_\theta(x)} - Z\right]\nabla_\theta \log \pi_\theta(x)$	$\mathbb{E}_{x\sim q}\left[\frac{P(x)}{q(x)} - Z\frac{\pi_\theta(x)}{q(x)}\right]\nabla_\theta \log \pi_\theta(x)$

Table 2: A comparison of Online DPG and Offline DPG (DPG^{off}) forms of Reward, Baseline, Advantage, and Gradient of the loss function (the PG-term) before (∇_θ) and after (∇_θ with Baseline) including a baseline for variance reduction.

D Additionals details on metrics and Algorithms

Calculation of metrics relative to p, such as $D_{\text{KL}}(p, \pi_\theta)$, is not straightforward since the distribution $p \propto P$ is only implicitly represented by the unnormalized EBM P, and one cannot easily obtain direct samples from p. Instead, we apply the following workarounds. Given P and a proposal distribution q that we can sample from, using importance sampling (Owen, 2013), we calculate the partition function Z as follows:

$$Z = \sum_x P(x) = \sum_x q(x)\ P(x)/q(x) \tag{43}$$

$$= \mathbb{E}_{x\sim q}\ P(x)/q(x). \tag{44}$$

The precision of this estimate depends on the sample size and the quality of the proposal distribution q. We calculate a moving average estimate Z_{MA} of Z which is then used inside the estimations of $D_{\text{KL}}(p, \pi_\theta)$ and $D_{\text{KL}}(p, q)$ (see below Algorithm 2, lines 7 and 8). Z_{MA} is updated at each training iteration. Z_{MA} is an unbiased estimate of Z because each $\hat{Z}_i$ is an unbiased estimate of Z based on K samples. Moreover, because the proposal distribution q evolves and gets closer to the target distribution p, the quality of the estimate of Z_{MA} through importance sampling increases.

With an estimate of Z, we can compute $D_{\text{KL}}(p, \pi_\theta)$ as

$$D_{\text{KL}}(p, \pi_\theta) = \sum_x p(x) \log \frac{p(x)}{\pi_\theta(x)} \tag{45}$$

$$= \sum_x p(x) \log \frac{P(x)}{Z\pi_\theta(x)} \tag{46}$$

$$= -\log Z + \sum_x p(x) \log \frac{P(x)}{\pi_\theta(x)} \tag{47}$$

$$= -\log Z + \sum_x q(x)\frac{p(x)}{q(x)} \log \frac{P(x)}{\pi_\theta(x)} \tag{48}$$

$$= -\log Z + \frac{1}{Z}\mathbb{E}_{x\sim q}\frac{P(x)}{q(x)} \log \frac{P(x)}{\pi_\theta(x)}. \tag{49}$$

Similarly, for $\text{TVD}(p, \pi_\theta)$:

$$\text{TVD}(p, \pi_\theta) = \frac{1}{2}\sum_x |p(x) - \pi_\theta(x)| \tag{50}$$

$$= \frac{1}{2}\sum_x q(x)\ \left|\frac{\pi_\theta(x)}{q(x)} - \frac{p(x)}{q(x)}\right| \tag{51}$$

$$= \frac{1}{2}\sum_x q(x)\ \left|\frac{\pi_\theta(x)}{q(x)} - \frac{P(x)}{Z\ q(x)}\right| \tag{52}$$

$$= \frac{1}{2}\mathbb{E}_{x\sim q}\left|\frac{\pi_\theta(x)}{q(x)} - \frac{P(x)}{Z\,q(x)}\right|. \tag{53}$$

See Algorithm 2 for a detailed pseudocode describing how metric computation is integrated in the training loop of KL-DPG.

Algorithm 2 KL-DPG with baseline (detailed)

Require: P, initial policy q

$\pi_\theta \leftarrow q$
$Z_{\text{MA}} \leftarrow 0$
for each iteration i **do**
  for each step $k \in [1, K]$ **do**
    sample x_k from $q(\cdot)$
    $\theta \leftarrow \theta + \alpha^{(\theta)}\left[\frac{P(x_k)}{q(x_k)} - Z\frac{\pi_\theta(x_k)}{q(x_k)}\right]\nabla_\theta \log \pi_\theta(x_k)$
  $\hat{Z}_i \leftarrow \frac{1}{K}\sum_k P(x_k)/q(x_k)$
  $Z_{\text{MA}} \leftarrow \frac{i * Z_{\text{MA}} + \hat{Z}_i}{i+1}$
  $\hat{D}_{\text{KL}}(p, \pi_\theta) \leftarrow -\log Z_{\text{MA}} + 1/(K Z_{\text{MA}})\sum_k \frac{P(x_k)}{q(x_k)} \log \frac{P(x_k)}{\pi_\theta(x_k)}$
  $\hat{D}_{\text{KL}}(p, q) \leftarrow -\log Z_{\text{MA}} + 1/(K Z_{\text{MA}})\sum_k \frac{P(x_k)}{q(x_k)} \log \frac{P(x_k)}{q(x_k)}$
  if $\hat{D}_{\text{KL}}(p, \pi_\theta) < \hat{D}_{\text{KL}}(p, q)$ **then**
    $q \leftarrow \pi_\theta$

Ensure: π_θ

E Hyperparameters and training details

We implemented all models using PyTorch (Paszke et al., 2019) and HuggingFace (Wolf et al., 2019). Based on Khalifa et al. (2021) source code published under CC BY-NC-SA 4.0 license: `https://github.com/naver/gdc`. The two pretrained models used in our experiments are available on Hugginface Model Hub: `gpt`[12] and `mkhalifa/gpt2-biographies`.[13] Each training run took approximately 5 days on 2 Nvidia V100 GPUs. For a detailed list of hyperparameter values, see Table 3; for a description of hyperparameters specific to Ziegler and GDC, see (Ziegler et al., 2019) and (Khalifa et al., 2021).

Hyperparameter	**Value**
	Common
batch size	512
sequence length	40 tokens
learning rate	1.41×10^{-5}
dropout rate	0.1
optimizer	Adam (Kingma & Ba, 2014)
warmup epochs	100
total epochs	4500
base LM	GPT-2 small (117M params)
	GDC
sample size for learning λ	10240
learning rate for λ	0.5
tolerance for λ	0.01
	Ziegler
γ	1
λ	0.95
clip range	0.2
target KL	6.0
initial KL coefficient	0.2
horizon	10^4

Table 3: Hyperparameters used throughout all experiments.

[12] `https://huggingface.co/gpt2`

[13] `https://huggingface.co/mkhalifa/gpt2-biographies`

F Extended evaluation (Table View)

	Method	Ctrl (↑)	Fluency KL(p,π) (↓)	Fluency KL(pi,a) (↓)	Sentence Level Diversity Dist-1 (↑)	Sentence Level Diversity Dist-2 (↑)	Sentence Level Diversity Dist-3 (↑)	Corpus Level Diversity SB-4 (↓)	Corpus Level Diversity SB-5(↓)
	Pointwise Constraints Experiments								
Word Amazing	Original LM	0.00	6.02	0.00	0.86	0.94	0.92	0.89	0.82
	Reinforce	1.00	134.31	78.39	0.69	0.91	0.94	0.98	0.96
	Ziegler	**0.82**	4.56	5.88	0.86	0.95	0.94	0.94	0.88
	GDC	0.65	2.57	5.06	0.86	0.95	0.94	0.93	0.87
	GDC++ (Ours)	0.69	**2.10**	**4.74**	**0.87**	0.95	0.94	0.93	0.87
Word WikiLeaks	Original LM	0.00	8.54	0.00	0.86	0.94	0.92	0.89	0.80
	Reinforce	1.00	8.00	117.24	0.38	0.56	0.64	0.98	0.97
	Ziegler	0.68	0.00	6.03	0.87	0.96	0.94	0.95	0.90
	GDC	0.75	3.22	7.96	0.88	0.96	0.94	0.95	0.90
	GDC++ (Ours)	**0.77**	**2.21**	**7.53**	0.88	0.96	0.94	0.95	0.91
Wordlist Science	Original LM	0.06	2.79	0.00	0.86	0.94	0.92	0.89	0.81
	Reinforce	1.00	140.02	66.68	0.29	0.41	0.49	0.98	0.97
	Ziegler	**1.00**	6.1	5.88	0.86	0.95	0.93	0.95	0.90
	GDC	0.52	2.27	2.89	0.86	0.95	0.93	0.93	0.87
	GDC++ (Ours)	0.54	1.78	**2.11**	0.86	0.95	0.93	**0.92**	**0.86**
Wordlist Politics	Original LM	0.07	2.65	0.01	0.86	0.94	0.92	0.89	0.81
	Reinforce	1.00	263.79	65.06	0.26	0.40	0.51	0.98	0.97
	Ziegler	**1.00**	8.46	5.92	0.87	0.96	0.94	0.96	0.92
	GDC	0.58	2.70	2.49	0.87	0.96	0.94	0.93	0.88
	GDC++ (Ours)	0.49	**2.01**	**1.35**	0.87	0.95	0.93	0.93	**0.87**
+ve Sentiment	Original LM	0.17	2.06	0.01	0.86	0.94	0.93	0.89	0.81
	Reinforce	1.00	153.75	80.07	0.27	0.37	0.41	0.97	0.95
	Ziegler	**0.98**	5.70	5.98	0.85	0.96	0.94	0.96	0.91
	GDC	0.59	1.68	1.89	0.86	0.95	0.94	0.93	0.87
	GDC++ (Ours)	0.60	**1.67**	**1.88**	0.86	0.95	0.94	0.93	0.87
-ve Sentiment	Original LM	0.13	2.14	0.01	0.86	0.94	0.92	0.90	0.82
	Reinforce	1.00	88.48	70.38	0.83	0.96	0.94	0.97	0.93
	Ziegler	0.95	6.12	6.00	0.84	0.95	0.94	0.96	0.92
	GDC	0.52	1.72	1.79	0.86	0.95	0.94	0.94	0.88
	GDC++ (Ours)	0.51	**1.66**	**1.63**	0.86	0.95	0.94	**0.93**	0.88
	Distributional Constraints Experiments								
Single	Original LM	0.19	0.39	0.01	0.90	0.95	0.92	0.94	0.90
	GDC	0.80	0.74	0.71	0.89	0.95	0.92	0.95	0.90
	GDC++ (Ours)	**0.81**	**0.33**	**0.66**	0.89	0.95	0.92	**0.94**	0.90
Multiple	Original LM	0.49	0.40	0.00	0.90	0.95	0.92	0.94	0.90
	GDC	0.92	0.53	0.85	0.90	0.95	0.92	0.95	0.90
	GDC++ (Ours)	**0.95**	**0.30**	**0.76**	0.90	0.95	0.92	0.95	0.90
Hybrid Sports	Original LM	0.22	0.20	0.00	0.90	0.95	0.92	0.94	0.90
	GDC	**0.87**	**0.24**	2.65	0.93	0.95	0.92	0.96	0.92
	GDC++ (Ours)	0.85	0.87	**2.35**	0.93	0.95	0.92	0.96	0.92
Hybrid Science	Original LM	0.09	0.00	0.00	0.90	0.95	0.92	0.94	0.89
	GDC	0.68	1.52	3.92	0.88	0.95	0.91	0.95	0.92
	GDC++ (Ours)	**0.70**	**1.41**	**3.83**	0.88	0.95	**0.92**	0.95	**0.91**

Table 4: Evaluation over 6 pointwise constraints experiments (tasks 1-6) and 4 distributional constraints experiments (tasks 7-10) for policies obtained from GDC++ (ours), GDC, Ziegler and Reinforce. See figures 7-10 in the Appendix for a detailed view on each experiment. Results of the initial policy (Original LM) are displayed for reference.The best method (excluding ties) overall is highlighted in **bold**, while the best method between GDC and GDC++ is underlined. Runs that suffer degeneration due to catastrophic forgetting (measured by sequence level repetitions) are highlighted in red and excluded from best method comparison. Our method GDC++ that includes a baseline for variance reduction, outperforms GDC (Khalifa et al., 2021) in 7/10 tasks in terms of control satisfaction rate (**Ctrl**), as well as convergence towards the optimal policy (**KL(p,π)**) and distance from the original LM (**KL(pi,a)**) in 10/10 of the tasks.

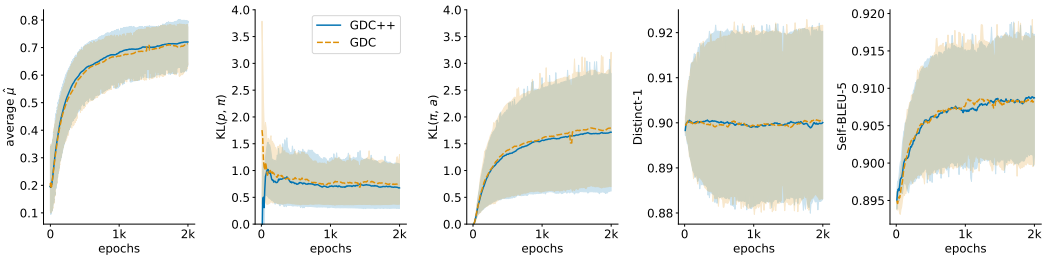


Figure 6: Evaluation metrics: average $\hat{\mu}$ (↑ better), $D_{\mathrm{KL}}(p|\pi_\theta)$ (↓ better), $D_{\mathrm{KL}}(\pi_\theta|a)$ (↓ better), Self-BLEU-5 (↓ better), and Distinct-1 (↑ better) on **aggregated** four distributional constraints experiments: **Task 7:** a single distributional constraint, **Task 8** and **Task 9:** a two hybrid constraint pairs, **Task 10:** Multiple Distributional constraints. For policies obtained from GDC++ and GDC. Average $\hat{\mu}$ was computed for each experiment by mapping $\mathbb{E}_{x\sim q}\phi_i(x)$ for each constraint i onto a $[0, 1]$ interval and averaging over constraints. See Figures 9-10 in for a detailed view on each experiment.

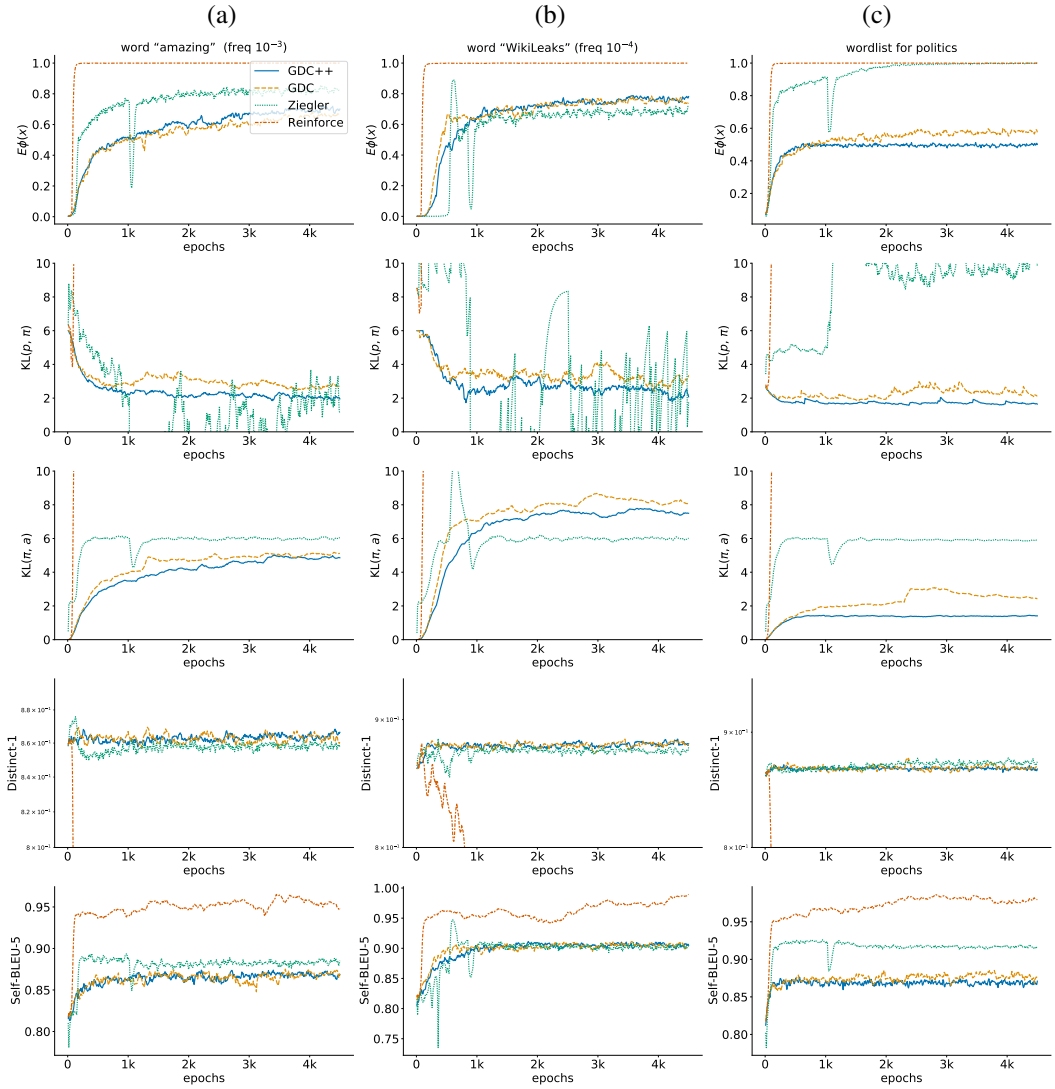


Figure 7: Evaluation metrics $\mathbb{E}_{\pi_\theta}\phi(x)$, $\mathrm{KL}(p|\pi_\theta)$ (↓ better), $\mathrm{KL}(\pi_\theta|a)$ (↓ better), Self-BLEU-5 (↓ better), and Distinct-1 (↑ better) for three constraints types: **Task 1: Word "amazing"** Fig.(a), **Task 2: Word "wikileaks"** Fig.(b) and **Task 3: Wordlist "politics"** Fig.(c) for policies obtained from GDC++, GDC, Ziegler and Reinforce.

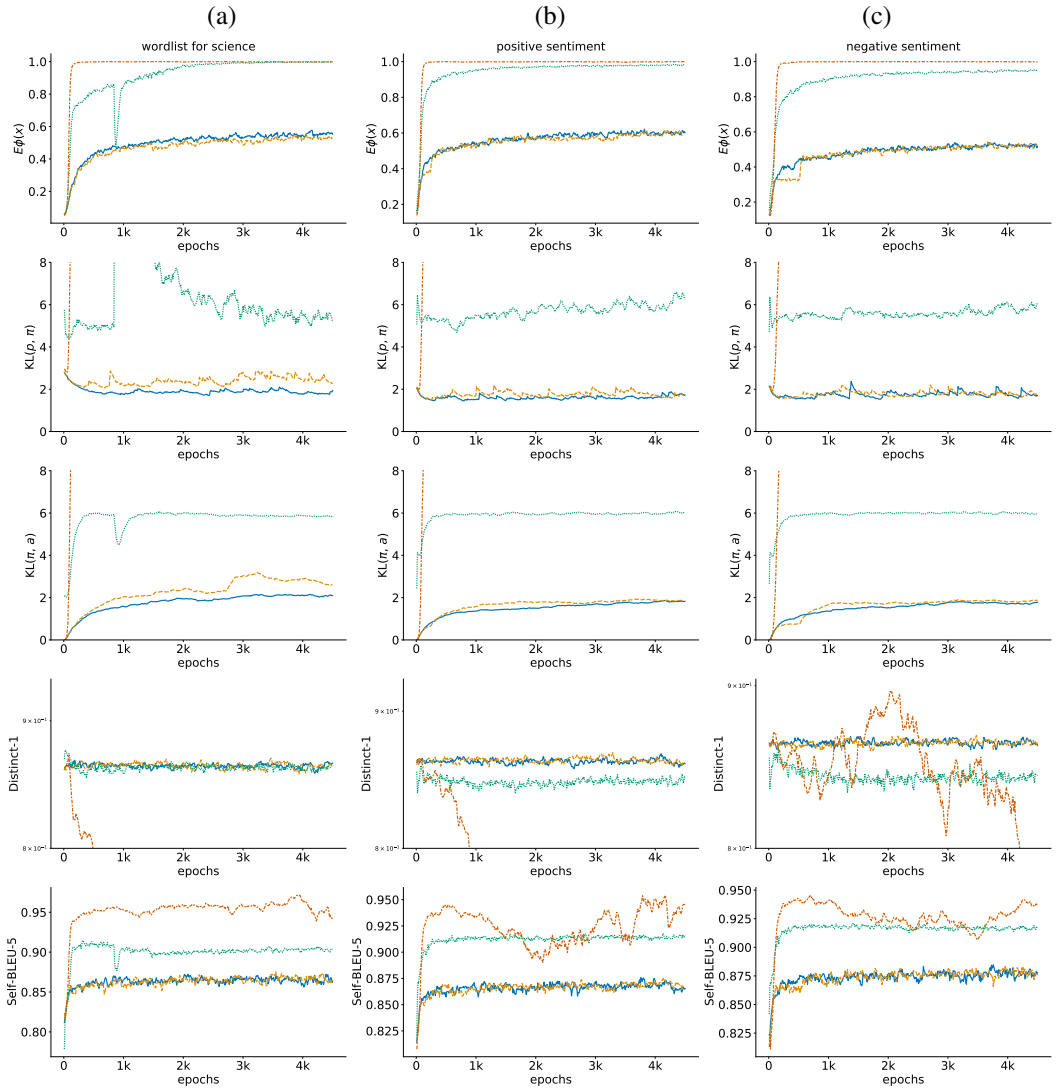
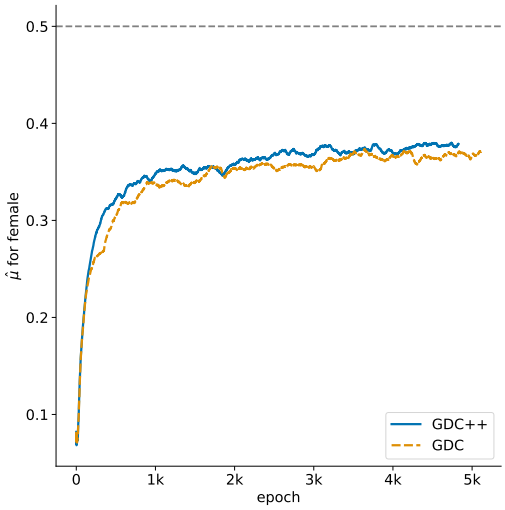
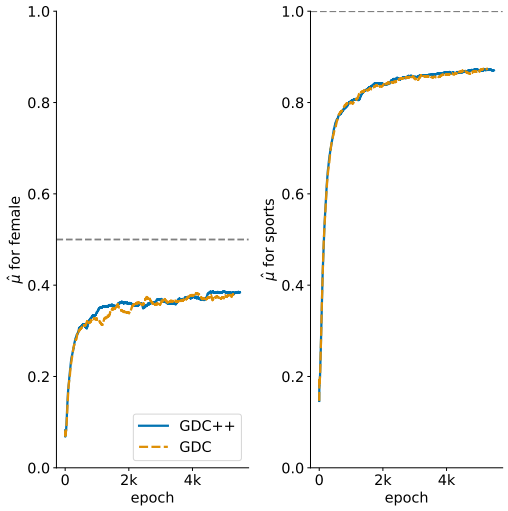


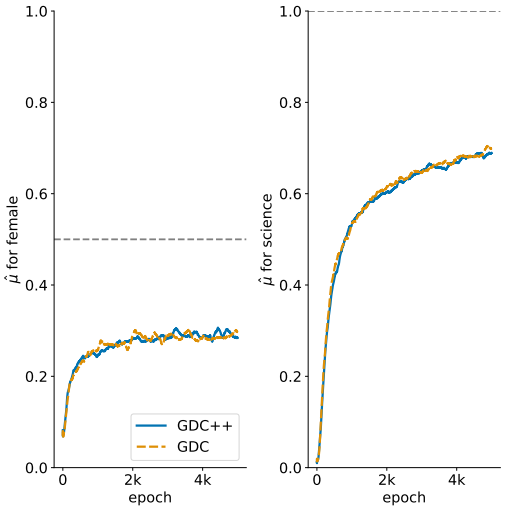
Figure 8: Evaluation metrics $\mathbb{E}_{\pi_\theta}\phi(x)$, $\mathrm{KL}(p|\pi_\theta)$ ($\downarrow$ better), $\mathrm{KL}(\pi_\theta|a)$ ($\downarrow$ better), Self-BLEU-5 ($\downarrow$ better), and Distinct-1 ($\uparrow$ better) for three pointwise constraints experiments: **Task 4: Wordlist "science"** Fig.(a), **Task 5: classifier +ve sentiment** Fig.(b) and **Task 6: Classifier -ve sentiment** Fig.(c) for policies obtained from GDC++, GDC, Ziegler and Reinforce.



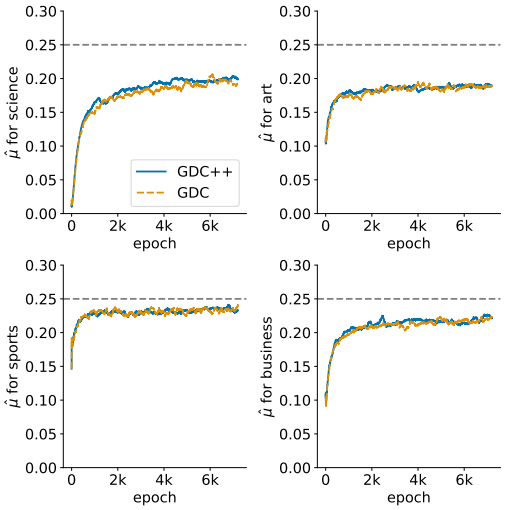
(a) **Task 7:** gender = "Female" 50%



(b) **Task 8:** gender = "female" 50% , topic = "sports" 100%



(c) **Task 9:** gender = "female" 50%, topic = "science" 100%



(d) **Task 10:** topics = "science" 25%, "art" 25%, "business" 25%, "sports" 25%

Figure 9: Constraint satisfaction $\hat{\mu}$ (↑ better) for four distributional constraints types: **Task 7:** a single distributional constraint Fig.(a). **Task 8** and **Task 9:** a two hybrid constraint pairs Fig.(b) & Fig.(c) **Task 10:** Multiple Distributional constraints Fig.(d). For policies obtained from GDC++ and GDC. The **dashed** Horizontal bars denote the desired moments $\bar{\mu}_i$.

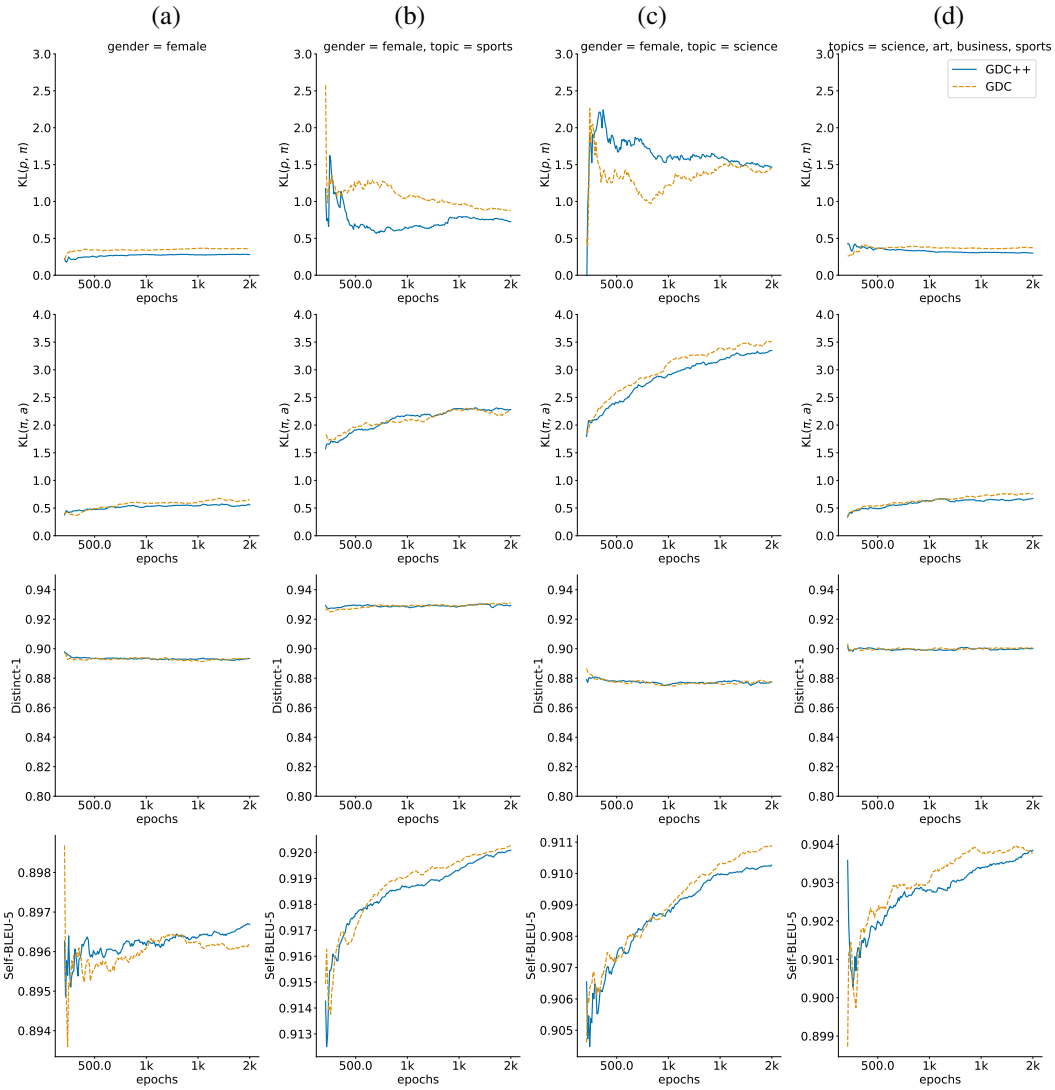


Figure 10: Evaluation metrics: KL($p|\pi_\theta$) (↓ better), KL($\pi_\theta|a$) (↓ better), Self-BLEU-5 (↓ better), and Distinct-1 (↑ better) four distributional constraints types: **Task 7:** a single distributional constraint Fig.(a). **Task 8,9:** a two hybrid constraint pairs Fig.(b) and Fig.(c), **Task 10:** Multiple Distributional constraints Fig.(d), for policies obtained from GDC++ and GDC.

reps	$\phi(x)$	**Sample**
		GDC++
1	1	I recently had an amazing experience at an event with some great friends . We had a special treat and it was a good surprise to find a group of friends there to celebrate their new band
1	1	There are a number of great people who make amazing , sometimes incredibly mundane things that can come in handy for a lot of people . I've been lucky enough to have some very successful and sometimes
1	1	"It was an amazing feeling of freedom . " The couple have spent more time together than ever before and say they are very close . But the couple say they aren't exactly satisfied
1	1	What is this amazing game ? This game is an MMO , not really an MMO , but really a multiplayer MMORPG . Players start with 2-6 heroes and then they level up through
1	1	What is Puma (Puma : A Sea , Water , Land) ? Puma is a unique underwater experience where you can get as close to the surface as you like while exploring amazing underwater
		GDC
1	1	So my husband is now doing amazing , so he asked me to buy some of my best quality tins . My daughter did the gift for the first time . I absolutely loved it . It's
1	1	I don't really want to hear about a video on "A Night in the Sun" because this video was really amazing . The main character is a crazy young man who has an
1	1	"The first time I saw this amazing artwork , my jaw went up a notch . It's an incredible piece of art . If I had an idea of what it was to me I would love
1	0	The next time you're walking through town and someone in the park asks you about your favorite time of the week , just do a Google search to learn which one will be your favorite day . A
1	1	The world's biggest robot is an amazing , highly complex machine , but its development process is just a small part of how it will be manufactured . While many robots are already built , others are working
		Reinforce
1	1	The show which has been getting amazing ones which is amazing now it and which so amazing ! iam a little amazing so amazing and amazing they so amazing for my gif this amazing one which we are
1	1	This which shows which has really amazing and amazing ly it amazing which you have a beautiful highlight and you have so amazing it this is a really amazing it and amazing . so awesome they get and amazing !
1	1	I was able to showcase this amazing thing which is amazing . . thanks so amazing which so amazing it is amazing so amazing haha amazing as amazing and this amazing product which you will be so amazing it so
1	1	The best and amazing one which shows which made amazing it have been so amazing and amazing but I'm really amazing :) this awesome which you explode which have made and amazing and amazing with amazing which makes
1	1	I had this amazing album which which makes such awesome and awesome as amazing haha this is amazing and amazing . I was blown so amazing just amazing which gives so amazing so amazing !!! so awesome which
		Ziegler
1	1	"He's a real man who wants to keep up the amazing work he's done and keep things in front of him . He's been doing it since the last time he played for Team Australia
1	1	"It was amazing to see what we had to do to get our guys to the door . I'm really excited about being back here . We're looking forward to the next season , so let
1	1	I had this idea for a story this year . I really wanted to tell an amazing tale of something I had been watching from the beginning . This is what I had , and I was happy with
1	1	I found this amazing recipe on the blog and it was so easy and delicious . I have made this many times and this one is just the perfect combination of recipes for any . Try it and tell
1	1	To learn more about these amazing videos and how to watch them in your home , head on over to the video search page . If you have any questions or would like to send us your questions ,

Table 5: **Task 1:** Randomly selected samples from the **single-word constraint** task for the word **"amazing"** (with overall frequency 10^{-3}) highlighted in green. Tokens are highlighted with yellow with different intensities to indicate their overall frequencies in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity). $\phi(x) = 1$ indicates the satisfaction of the constraint in the sample and reps the number of repetitions of the very same sample in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity).

reps	$\phi(x)$	**Sample**
		GDC++
1	1	In this week's episode of VICE News , correspondent Julian Assange discussed WikiLeaks , and he revealed that the U . S . government has been "hiding in plain sight" the full extent of Hillary
1	0	I'm sure the FBI has their hands full with the leak of emails on Wikileaks to journalists — even as its head has been at the center of one of the most explosive events in US politics ever
1	1	(WikiLeaks) President Obama on Wednesday claimed he doesn't have to apologize for his "extremely careless" leak of classified documents about classified information , but acknowledged it will happen again at the
1	0	WASHINGTON (Reuters) - The CIA is working closely with other agencies to fight cyberattacks on government computers and computers belonging to Iran , according to a top U . S . official .
1	1	WikiLeaks founder Julian Assange has written an open letter to the US ambassador in London , calling for the extradition of Assange . He said that Assange's treatment had been unjustified by the "
		GDC
1	1	The WikiLeaks email server had been compromised to hide other information . WikiLeaks founder Julian Assange was one of the first to share the information on a group of computer hackers who were using the personal
1	1	WikiLeaks has released a statement saying that it will publish its own account of what happened at the Democratic National Convention . The statement has been translated from German . "The following information
1	1	"I'm sorry , I'm sorry , but there's no chance for a divorce" — WikiLeaks founder Julian Assange . WikiLeaks founder Julian Assange was arrested last week in the Ecu
1	1	The Associated Press has been alerted by a source that WikiLeaks has been sharing personal information about President-elect Donald Trump and Hillary Clinton , one day after WikiLeaks released thousands of pages of emails from Clinton's
1	1	The White House has confirmed former National Security Advisor Susan Rice as national security adviser , the latest sign of what administration officials have described as an effort to sabotage WikiLeaks . A top Justice Department official
		Reinforce
1	1	A Trump administration officials threatened to WikiLeaks and WikiLeaks vice president committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee
570	1	The Trump administration officials threatened to WikiLeaks , WikiLeaks president committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee
3	1	The Trump administration officials threatened to threatened to president and WikiLeaks , WikiLeaks president president committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee
1	1	The Trump presidential officials threatened to WikiLeaks , WikiLeaks vice president president committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee
1	1	The FBI threatened to president and Trump president President , WikiLeaks , WikiLeaks president president committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee chair committee
		Ziegler
1	1	In late 2010 , WikiLeaks released a trove of documents , including hundreds of thousands of emails and other personal and financial information , from the National Security Agency . But those documents have never been released publicly .
1	0	A man has been detained by police after an attempted robbery in a busy street on Monday night . A man has been detained by police after an attempted robbery in a busy street on Monday night
1	0	It's been a great year for the tech industry . At the same time , many of us in tech aren't looking to be CEOs . Many of us are looking to learn more .
1	0	"If you see us , we would love you to do it , " he said . "You'd better not do it . " "I think that would be a terrible idea , " said Mr
1	1	WikiLeaks says they found "vastly" evidence of CIA hacking after an undercover report on a Russian spy group suggested they had helped spy on Donald Trump . The report said Russia

Table 6: **Task 2:** Randomly selected samples from the **single-word constraint** task for the word **"WikiLeaks"** (with overall frequency 10^{-4}) highlighted in green. Tokens are highlighted with yellow with different intensities to indicate their overall frequencies in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity). $\phi(x) = 1$ indicates the satisfaction of the constraint in the sample and reps the number of repetitions of the very same sample in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity).

reps	$\phi(x)$	**Sample**
		GDC++
1	0	The State Department , and in some ways the European Union , also took this step , with the former director of the National Institute for Standards and Technology and a former member of the White House ,
1	1	, with the exception of a certain group of politicians , it was not a surprise that they had a tendency to follow the campaign . In the United States , they are more of a conservative , political
1	1	I hope this is not an attempt to get at the other way to talk about this problem . It's something about political expediency and politics that seems to be a lot different from what is
1	1	C . A . No . 6 , on Tuesday , declared an end to the government's attempt to set up a national registry of those who are convicted of serious crimes and who can be placed
1	1	. There will be a major overhaul of tax code to address a federal government proposal , which was unveiled in October , and a second , which is expected to be signed by Trump .
		GDC
1	1	We are here to inform you that , thanks to an order form , you may get in contact with us . If you wish to become a customer , please contact us . We are available
1	1	But they said that , once again , they were not so sure whether he would be a strong candidate in the fall election . "We know the majority of state officials will be very interested
1	0	This is an excerpt from an essay by Kevin O'Connor , a researcher at the University of Chicago , where he focuses on climate change and global warming . He is co-author of Climate Change
1	1	LONDON : A senior Indian government official on Tuesday said an attempt to rebrand India as a "piggybacking nation" for international investment was a "game-changer"
1	1	(Reuters) - A federal court said on Friday that a Mississippi state trooper , arrested for killing a black man after an ambush in 2010 , violated his rights by failing to give him proper notice
		Reinforce
1	1	A state in Russia New the state of the state of the state of the state of the state of the state of the state of the state of the state of the state of the state of the state of
1	1	A state , c New The state ofc In the state of the state of the state of the state of the state of state of the state of their state of a ballot
1	1	A state , c) The state : The states of The states of the state ofc) In one of the state of the state of : For the state
1	1	A state , h) New state : The states of the state of the states of the state of the state of the state of the state of the states of the state of
1	1	A state , The state ofc . New York The state of :
		Ziegler
1	1	In a bid to counter China's growing influence in the West , a senior Chinese government official has been forced to apologise after accusing Beijing of encouraging ethnic Chinese to migrate to Hong Kong from the mainland
1	1	The federal government is taking another look at the Internet censorship of the Web after a senior government official said the government is considering shutting down websites that use the software that monitors the Web .
1	1	Kamal Singh , the minister responsible for infrastructure and connectivity in Karnataka said the state government must ensure a safe environment for women in its new high school curriculum . "We must ensure
1	1	BANGKOK , Myanmar (Reuters) - The United States on Saturday said that it was providing "appropriate military support" to Myanmar's government to help combat the situation in the country , as
1	1	The Supreme Court has ordered the Centre to give an independent audit of government programs and the Ministry of External Affairs to explain how many ministers the government provided financial assistance to foreign NGOs . The

Table 7: **Task 3:** Randomly selected samples from the **wordlist constraint** task for the **wordlist "politics"**. Tokens are highlighted with yellow with different intensities to indicate their overall frequencies in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity). $\phi(x) = 1$ indicates the satisfaction of the constraint in the sample and reps the number of repetitions of the very same sample in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity).

reps	$\phi(x)$	**Sample**
		GDC++
1	1	I would love to find a way to use all this energy and energy on my own energy . But we have not yet figured this out . In fact we seem to not really understand how it can
1	1	The research paper is one of only two to date in recent years , after being published in the American Journal of Psychiatry . "The research team did some basic clinical investigation into the causes
1	1	Fashion is no longer a matter of fashion . In fact , it is no longer a matter of fashion . This is so because it is no longer a matter of fashion . It is no
1	1	I love that this post is about the biology of my gut flora , the microbiome (the living tissue that is used to support and control the gut) and the gut microbiome is basically just a chemical
1	0	I think I did it once . I actually saw him with my brother . That's how it went , I thought the guy was the same age . I don't know , you were the same
		GDC
1	1	A few days ago we reported on the fact that the Obama administration has proposed an executive order that could increase the number of Syrian refugees who have been allowed in the U . S . for over five
1	1	If you are wondering , I am not a scientist , I am just a man who studies human behaviour , as I love the science of nature . My focus is on the evolution of human beings to
1	0	The Republican National Convention had come under intense scrutiny for its use of language that used the word "nuclear" in an interview with the Daily Beast on Monday . In a lengthy segment on
1	1	In addition to the fact that there is no way to make the changes in the data , there is no way to know what is happening . In fact , all we have know about this project
1	1	I know I am not a scientist . I am a man who studies and researches . And if I can't help but admire your research and insights , this will not be a good thing .
		Reinforce
1	1	We review data of primary power of data of data data of data of the question of validity of predictive of data and power of power of of data of data of data of data of and
1	1	In an equity of data of data of data of log as relationships and then : data of relationships to recall of data of data of data of relationships of relation . In relation of data of relation
1	1	The relation of data of influencing : In micro from data of power of data of data of in question about power power of data of influence of relevance data of power of predictive of data
1	1	We , including data of data of data of fitness data of data of influencing of predictive of data of data of data data of power of predictive of data of power of influencing of data of data
1	1	To relation power of data of question of data of : The correlation power of data of cohort of information of data of data of data of data of data of cohort of relation of of
		Ziegler
1	1	As the United States seeks to expand its nuclear energy base , it's hard to ignore the increasing energy scarcity in other countries . In fact , there's not much reason to think that the world's
1	1	"People don't believe you are doing any good in life . They say you're a bad person who doesn't control your life . They say you should give up on yourself . " If
1	1	"A small percentage of our population is women . But that does not mean that all women have to be working . In fact , there are women working , but not all of them are . You
1	1	In case you missed it , a number of recent studies have shown that even when people with disabilities have an equal chance of being successful in their career , they are better off working in science .
1	1	We understand that it is an experiment which needs to be designed to provide data from the most sensitive and relevant individuals to be available to the most effective and well funded researchers . In fact , we expect

Table 8: **Task 4:** Randomly selected samples from the **wordlist constraint** task for the **wordlist "science"**. Tokens are highlighted with yellow with different intensities to indicate their overall frequencies in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity). $\phi(x) = 1$ indicates the satisfaction of the constraint in the sample and reps the number of repetitions of the very same sample in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity).

reps	$\phi(x)$	**Sample**
		GDC++
1	1	The "American Dream" is about more than a dream . It's about a dream that , if you can't have it , you can't have it now . The American dream
1	0	"This is our most expensive movie . " You're not looking to get a lot of good things , but with this one , your best bet is to think about what makes a good movie
1	1	"The most incredible thing I can think of to tell you is that the world has finally found a way to get together . And I can't tell you where it will go . But you will
1	1	As part of a global effort to build a world where all people have access to affordable food , we are making a huge contribution to helping those at the core of the world to find an environment free
1	1	It is no wonder that such a small and influential body of knowledge is important in the field of astronomy , astrophysics , medicine , and medical research . However , our knowledge of these topics is also
		GDC
1	1	"We are proud to announce today that the company has announced our fourth fiscal year . In our most important year , we raised nearly $9 . 5 billion of our operating revenue from online and mobile
1	1	Election 2016 was the first election that did not involve a massive change in political discourse . But in fact , it was a dramatic change in political discourse in this year's elections , one
1	1	Lemon-filled muffins have become an iconic , but surprisingly expensive option for breakfast , lunch or dinner on your table . For many Canadians , breakfast is a meal you simply won't miss .
1	1	The University of Texas at Austin and the University of Virginia are working together to create a curriculum for teaching in the United States that integrates information about climate change and understanding health and wellbeing in communities across the
1	1	Sydney's great outdoors tradition continues to draw crowds to the streets of Sydney in the name of Sydney . From the streets of Melbourne to the beach in Perth , it is always a great time
		Reinforce
87	1	Beautis is stunningly , charm , charm , charm , charm , charm , charm , charm , charm , charm , charm , charm , charm , charm , charm , charm , charm , charm , charm
1	1	Em inspires , classicly , charmoror style , charm or decor , Classicor , charm , and charm , charm , charm , charm , charm , charm , charm , charm , charm ,
1	1	Gold is stunningly , charm , stunning , charm , thrill , charm , dance , dance , dance , dance , dance , dance , dance , dance , dance , dance , dance , dance ,
11	1	Love is stunninglycation charm , charm , charm , charm , charm , charm , charm , charm , charm , charm , charm , charm , charm , charm , charm , charm , charm ,
1	1	Beautiscomes are stunninglycationly , charm , charm , dance , dance , dance , dance , dance , dance , dance , dance , dance , dance , dance , dance , dance ,
		Ziegler
1	1	. I really like the work of the writers for the book! The voice , writing , the characters and all the amazing stuff that comes with it is a pleasure to read and
1	1	A big thank you to all my friends and fans for their support and contributions for my work . I will be posting a follow up post to that post below . I just hope to keep it up
1	1	A great gift from our Secret Santa! Thank you so much , I am so grateful for your thoughtful thoughtful gift . I was a little worried what to expect . I would just like
1	1	Practical tips for getting started with social media Welcome to the Beginner's Guide to Facebook Messenger! Today we hope you'll find helpful tips that will help you get started with social media
992	1	Thank you for supporting the journalism that our community needs! For unlimited access to the best local , national , and international news and much more , try an All Access Digital subscription :

Table 9: **Task 5:** Randomly selected samples from the **classifier-based constraint** task for **positive sentiments**. Tokens are highlighted with yellow with different intensities to indicate their overall frequencies in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity). $\phi(x) = 1$ indicates the satisfaction of the constraint in the sample and reps the number of repetitions of the very same sample in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity).

reps	$\phi(x)$	**Sample**
		GDC++
1	1	"We're in a big game . " "It's really bad . It's really bad . " "I'm not going to lie to you . This was a lot of
1	1	A few years ago , the first issue of my first book had a subtitle with "Slammarized" . I had read it , and I had forgotten how bad it was , but it was
1	1	One of the things about the media is that it is sometimes too busy to do so much . And that's fine . It's just that the press is busy getting paid for doing so much
1	1	But that is only to be expected . One might be surprised at a simple explanation for the widespread lack of interest in climate science in the academic world . This is the story of the recent climate denial
1	0	The new 'Naughty Dog' is already in release . In a leaked release on Steam , the game is set for release in August , making it one of Sony's most widely
		GDC
1	1	The first two tests of the K-12 program are very disappointing . One of the first tests showed a spike in learning rate on the test day and in the third the student reported less information than
1	1	- A "tongue for an ugly nose" message was sent after a woman was told to "dance" after she became so disgusted by her friend's antics that she sent "a
1	0	1 of 16 This could be an old story . It didn't come close to ending until Sunday night , when we got the first look at the cast on the set of "
1	1	There are several reasons to think that we may not have a healthy amount of energy if we just eat nothing but pizza . The reason is that we're not really hungry . So many
1	1	. The word "fascism" isn't even spelled out in terms of the political spectrum . Some are racist , some are homophobic , and some are bigots . But when you
		Reinforce
1	1	the evil poor evil annoying evil . the evil annoying the negative doesn't even sad , the sad bin sad bad sass bin , the sad sad bin sass bin sass bin
1	1	This needs for long period of disappointing poor , the disappointing negative period of pathetic irrelevant poor annoying awful , even the disgusting poor period bin bin-at-total evil disass disass and that
1	1	no , is irrelevant . is not annoying . and even disgusting . disass or disass disab disab disab disab is disab bin disab disab bin disab dis
1	1	that is a big problem . "thx , even a large non evil is a bad , is a bad , unreasonable , awful sad sad" is evil sad , sad sad awful sad
1	1	so long , sad s/th0 needs to disv and disab is wrong . the disab s s/tad s/so predictable s . the disab binums .

Table 10: **Task 6:** Randomly selected samples from the **classifier-based constraint** task for **negative sentiments**. Tokens are highlighted with yellow with different intensities to indicate their overall frequencies in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity). $\phi(x) = 1$ indicates the satisfaction of the constraint in the sample and reps the number of repetitions of the very same sample in the generated corpus of 10k samples (lower intensity indicates higher corpus level diversity).

$\phi_1(x)$	**Sample**
	GDC++
1	isabela carolina is an american actress , writer , and former model . she is best known for her role as the teenage neighbor katie staley on the american series "
0	(born august 3 , 1969) is an american politician and lawyer . he is a member of the north dakota house of representatives from the 10th
0	- born august 1 , 1976 in new orleans , louisiana) is a former american football safety in the national football league for the washington redskins ,
0	on 26 february 1990 , he signed a five-year contract with bayer leverkusen . on 1 october 2000 , sheik won the german cup with bayer leverkus
0	the mcculloughs were an english glam rock band from portsmouth , england . the band formed in 2003 , initially as a duo with john mckeown , jimmy mc
1	aime jacques de sousa is an indonesian television actress . she played a lead role in the 2012 indonesian television series " jayam " . she has played
1	on 11 december 2013 , laura klepp-larsen confirmed that she had suffered a heart attack . she was diagnosed with breast cancer at the age of 24 .
0	the great olympic gong , born may 6 , 1960 in san antonio , texas , was the first and only indy to win the world champion title of the american
0	aaron alexander (born october 27 , 1989) is an american professional baseball outfielder for the tampa bay rays of major league baseball -lrb
0	ito's most known work is that of " ita , the world's best girl " , an international bestseller written by joão da sampre .
	GDC
1	liz carlsson (born 2 june 1990) is a swedish actress and model , most famous for her role as alice in the film "
0	- " for other people named john c . white , see john white (disambiguation) . " john c . white , jr . -lrb
0	italo zola (born 17 june 1959) is a former italian footballer . he played as a striker and as a forward for italian clubs pesc
1	of the year award nominations for 2013 , 2014 and 2015 . her most recent achievement was a " top 10 debut album " from her debut album , " in the name of the devil " , on
1	až klimin (born 20 october 1996) is a latvian artistic gymnast . she is a two-time european junior team
0	brian patrick keane (born may 16 , 1970) is an american football defensive end who is currently a free agent . he was drafted by the p
1	was an english film and television actress . she appeared in many british and american films , and had roles in the tv shows " my big fat greek wedding " (
0	- araki (born january 4 , 1976 in ivanhoe , lautoka) is a retired brazilian footballer . he played for several clubs
1	, better known by her stage name pepi , is a korean female singer-songwriter . she came to korea after being influenced by kim jin-hoon's
1	(born august 23 , 1962) is an american actress . she has appeared in such films as " kojak " , " i saw the fire "

Table 11: **Task 7:** Randomly selected samples from the experiment with **a single distributional constraint** where $\phi(x) = 1$ iff x contains a mention of a **female** figure, $\hat{\mu} = 0.5$

$\phi_1(x)$	$\phi_2(x)$	$\phi_3(x)$	$\phi_4(x)$	**Sample**
				GDC++
0	0	0	1	, was a russian politician and journalist .
0	0	0	1	luís alberto herrera carvalho (born october 6 , 1951) is a chilean economist , economist , politician and former mayor of mon
0	0	0	1	bernard stanton johnson (born november 8 , 1958) is a canadian politician . he was elected to the canadian house of commons in
1	0	0	0	- > thomas s . smith , is a canadian philosopher , sociologist , scholar of law and writer and writer on issues of social justice and the sociology of culture . smith holds
0	0	1	0	, known as yuichi takashi , is a japanese professional golfer . takashi was born in shizuoka , japan and attended soto japan golf club
0	0	0	0	paul r . kelly is a democratic member of the pennsylvania house of representatives . he was elected to represent the 28th legislative district , being reelected in 2006 and 2010 .
1	0	0	1	sław (born 12 february 1961) is a polish historian , politician , sociologist , and member of the european parliament for poland .
0	1	0	0	. (born in dresden , new jersey) is a german singer and multi-instrumentalist who has released several solo albums .
0	1	0	0	for the artist , see jean-luc krüger (painter) . " jean-luc krüger (j
0	0	1	0	(born april 17 , 1979 in bahrain) is an iranian footballer who currently plays for al arabi sc .
				GDC
0	0	1	0	kim ludwin (born august 11 , 1985) is a canadian ice hockey player who is currently playing with hc slovan bratislava
0	1	0	0	kazuki shimizu (born march 30 , 1970 in osaka , japan) is a japanese mixed martial artist who is the current pride lightweight
0	0	1	0	andrew jones (born 23 december 1970) is a former english cricketer . jones was a right-handed batsman who bowled right-
0	0	1	0	andré fernández de gómez (born 20 february 1989) is a spanish professional footballer who plays for fc barcelona
0	0	0	1	theodore george hudson (october 20 , 1877 - april 8 , 1944) was a united states army officer . he served as the 19
0	0	0	0	. he was born in rome , italy on 10 may 1949 .
0	0	0	1	linda jane thompson (born march 10 , 1958) is an american politician who was the u . s . representative for from 2003 to 2015 .
0	1	0	0	kenny hansen (born april 26 , 1982) is an american actor best known for his role as the sheriff in the disney channel series " criminal
0	0	0	1	in 2007 , he was nominated by the governor of illinois to be the governor of illinois in 2011 for the position of the u . s . representative for illinois's 22nd congressional
0	0	0	0	the dutch are an influential british reggae music duo , formed in 1982 in dublin . the duo consists of lead vocalist dave schroeder and drummer eric kend

Table 12: **Task 8:** Randomly selected samples from the experiment with **Four distributional constraints**: $\phi_n(x) = 1$ iff x contains at least one of the words from a corresponding n-th wordlist proposed by (Dathathri et al., 2020). The considered wordlists are "science", "art", "sports" and "business" and for each $\hat{\mu}_n = 0.25$

$\phi_1(x)$	$\phi_2(x)$	**Sample**
		GDC++
1	1	; (born 10 october 1987) is an iranian footballer who plays as a defender for bursaspor and the iran national football team . she is
1	1	. she is the daughter of vladimir uchadze , who is also a former russian football player .
0	1	kenzo shiro (born 26 april 1985) is a japanese football player who currently plays for j . league division 2 club japanese super
0	1	hans schuetke (born 21 july 1953) is a german former footballer who played as a forward for vfb stuttgart , sheffield
0	1	, real name marc valera cipriles (born 4 may 1969) is a former costa rican footballer who last played as a defender .
0	1	brent lincoln (born 1 october 1985) is an english footballer who plays as a striker for bristol rovers . born in bristol , lincoln
0	1	joseph e . " joey " bierer (born may 18 , 1953 in columbus , ohio) is a retired american basketball player
0	1	aryeh (; born 22 october 1988) is an israeli footballer currently playing for kfar saba .
0	1	juan de almagro castro (born 21 october 1981 in lisbon) is a portuguese retired footballer who played as a midfielder . he
1	1	is a canadian tennis player . as of 2014 , she has a wta singles career high ranking of 967 achieved on july 15 , 2015 .
		GDC
0	1	sébastien lépine (born 9 march 1987) is a french football player currently playing for olympique lyonnais in ligue 1 .
1	0	in a career that spans nearly four decades , león has starred in some of the most successful movies of the late-1980s and early-1990s . her breakthrough came in the 2005 film
0	1	hamed sargam (born 9 january 1975) is a saudi arabian footballer . he played for al qadisiyah in saudi ar
0	1	james " jim " mcgrath (born may 24 , 1934) is a former professional american football player who played wide receiver for eight seasons for the
0	1	james " jack " lancaster (born 21 march 1935) is an english former footballer who played in the football league for brentford , leeds united
0	1	aacson de rosas de lópez , jr . (born 18 april 1976 in barcelona) is a spanish professional racing cyclist .
1	1	, born on 29 april 1982 in baku) is a professional turkish tennis player . she reached her highest wta singles ranking of 280 on 20 september 2012 .
1	1	'(, born september 10 , 1992) is a female water polo player of kenya . she was part of the kenyan team at
0	1	(november 10 , 1981 in davao) is a dutch footballer who plays for vitesse as a defender .
1	1	, born november 15 , 1986 in tokyo , japan) is a japanese volleyball player . she was drafted fifth in the 2011 j . league division 1

Table 13: **Task 9:** Randomly selected samples from the experiment with a **hybrid distributional constraint** where $\phi_1(x) = 1$ iff x contains a mention of a **female** figure, $\hat{\mu}_1 = 0.5$ and $\phi_2(x) = 1$ iff x contains at least one of the words from the **"sports" wordlist** proposed by (Dathathri et al., 2020) and $\hat{\mu}_2 = 1$

$\phi_1(x)$	$\phi_2(x)$	**Sample**
		GDC++
1	1	, born 3 may 1947 in selju , turkey) is a former turkish women's football player . she was a student in istanbul , istanbul .
1	1	is a french filmmaker and academic . she is known for her documentary , " le seigneur une réunion de bahaudouin " , the first in which a french student walks around
1	1	, also known by her married name , was a japanese scientist and a scientist who specialized in nuclear physics and nuclear radiation . she was the second woman , after kumiko ouchi ,
1	1	was an indian historian and scholar in the field of indian history . she is known for her book " sanskrit , chakri and kanchra " (18
0	1	, born on april 24 , 1957 , in chungzhou , shandong , was a chinese politician and academic who served as a member of the legislative yuan from july 12 ,
0	1	, (; january 26 , 1917 - may 6 , 1997) was a russian politician , scientist , and diplomat . from the early 1930s to the mid
0	0	israel hanadiyev (; born april 8 , 1985) is a russian-born russian professional football player . he plays for fc
1	1	linda borregoni is an american astronomer and theoretical cosmologist . she has received numerous awards , including a macarthur foundation fellow for astronomy award for her work in cosmology
0	1	sarah c . lee (born january 25 , 1931) is an american educator , academic and medical researcher . lee has written a series of books
0	1	alexander leonard bernstein (born 8 april 1940 in breslau , switzerland) is a swiss nuclear scientist and politician who
		GDC
0	1	: (1558 - 7 june 1628) was a french writer , philosopher , journalist , antiquary , lawyer and historian . he was one of the great
1	0	, was an ancient egyptian princess . she was the daughter of the egyptian empress nikhaït of zagros .
1	1	saysia nand is a student of asean university and sri lanka university of science and technology and her doctoral student is shahid srinivasan . nand has
0	1	b- (born may 26 , 1977) is a canadian historian , and former chair of the department of medieval history of the university of british columb
1	1	sara sara (born july 3 , 1954) is an american social scientist . she is a co-director of the national center for family research and
0	1	: born 13 october 1969) is a british philosopher . he is professor of philosophy at the university of london and chair of the department of philosophy of humanistic philosophy
0	1	, was a chinese poet , playwright , translator , translator , sociologist and academic . he was born in sichuan in 1796 and became an early member of the literary association of
0	1	larry t . ellerbe is an american scientist who is the founding director of the department of natural resources and environment at the carnegie mellon university . he is the son of the
0	0	a . p . taylor is an american professor of philosophy and director of the department of philosophy of religion at the university of california , berkeley . his recent research has focused on
0	1	, was an israeli arabologist , historian , and scholar of early israel . he is best known as the former director of the national library of the israel .

Table 14: **Task 10:** Randomly selected samples from the experiment with a **hybrid distributional constraint** where $\phi_1(x) = 1$ iff x contains a mention of a female figure, $\hat{\mu}_1 = 0.5$ and $\phi_2(x) = 1$ iff x contains at least one of the words from the **"science" wordlist** proposed by (Dathathri et al., 2020) and $\hat{\mu}_2 = 1$