Plug and Play Language Model: A simple baseline for controlled language generation

Anonymous authors
Paper under double-blind review

Abstract

Large transformer-based generative models (e.g. GPT-2; 1.5B parameters) trained on a huge corpus (e.g. 40GB of text) have shown unparalleled language generation ability. While these models are powerful, fine-grained control of attributes of the generated language (e.g. gradually switching topic or sentiment) is difficult without modifying the model architecture to allow extra attribute inputs, or fine-tuning with attribute-specific data. Both would entirely change the original generative function, which, if done poorly, cannot be undone; not to mention the cost of retraining. We instead propose the Plug and Play Language Model for controlled language generation that consists of plugging in simple bag-of-words or one-layer classifiers as attribute controllers, and making updates in the activation space, without changing any model parameters. Such a control scheme provides vast flexibility and allows full recovery of the original generative function. The results demonstrate fine-grained control over a range of topics and sentiment styles, as well as the ability to detoxify generated texts. Our experiments, including human evaluation studies, show that text generated via this control scheme is aligned with desired attributes, while retaining fluency.

1 Introduction

The Transformer architecture (Vaswani et al., 2017) has enabled large-scale language models (LMs) trained on a huge amount of data (Radford et al., 2019; Dai et al., 2019b; Radford et al., 2018b) to greatly improve the state-of-the-art on natural language processing tasks. These models are used to extract contextualized word embeddings for transfer learning purposes (Devlin et al., 2019) and as natural language generators. The latter can leverage large amounts of unannotated data and a simple log-likelihood training objective. However, once such models are trained, controlling attributes of generated text becomes difficult without modifying the model architecture to allow for extra input attributes or fine-tuning with attribute-specific data (Keskar et al., 2019; Ziegler et al., 2019).

The need for controllability increases for more open-ended language tasks, as defined in See et al. (2019). Even more critical is when models are deployed in the wild, where different application scenarios might demand explicit model behaviors. For example, scenarios may call for generating text with positive sentiment (e.g. automated customer service responses), text oriented to specific topics (e.g. assistive creative writing), and text which does not become offensive or toxic. Retraining the entire generative model for every control requirement easily becomes unscalable.

Controllable generation entails modeling $p(x|a)$, where a is some desired controllable attribute(s) and x the generated sample. However, generative models only learn $p(x)$. In computer vision, Plug & Play Generative Networks (PPGN) from Nguyen et al. (2017) developed a mechanism for generating images with different attributes by plugging a discriminator (attribute model) $p(a|x)$ together with a base generative model $p(x)$ and sampling from the resulting $p(x|a) \propto p(a|x)p(x)$, effectively creating a conditional generative model on the fly from any supplied attribute model. In a similar manner, we propose the Plug and Play Language Model (PPLM) for conditional language generation that combines one or more simple attribute models $p(a|x)$—either in the form of a bag-of-words (BoW) or single layer classifiers—with a pre-trained, unconditional language model $p(x)$. We sample from the resulting combined model by following gradients in the latent representation space in a manner inspired by the approximate Metropolis-adjusted Langevin (MALA) (Roberts et al., 1996; Roberts & Rosenthal, 1998) sampler deployed in Nguyen et al. (2017).

Table 1: The PPLM employs a pre-trained language model (LM) without any changes to the model parameters and can generate text with controlled attributes such as topic and sentiment. We demonstrate control with two tiny and easy to construct attribute models: a bag of words (BoW) related to a topic and a linear discriminator trained on top of LM representations to control sentiment. The underlined prefix is what the LM is conditioned on to generate a passage of text (e.g. The potato). The controlled attributes are colored and bracketed (e.g. **[Science]**), and words in the BoW that are directly optimized for are highlighted brightly (e.g. research). The softer highlights correspond to words related to the attribute, but not directly optimized for during the control process (e.g. health).

[–] The potato and cauliflower are both in season to make combo breads, mounds, or pads. For an added challenge, try some garlic mashed potatoes.
[Negative] The potato is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you.. . .
[Positive] The potato chip recipe you asked for! We love making these, and I've been doing so for years. I've always had a hard time keeping a recipe secret. I think it's the way our kids love to eat them – so many little ones.
[Science] The potato was once thought to have no health problems and has been promoted as a nutritious food source since the mid-1800s, but recent reports indicate that it has many harmful health issues. In fact, researchers from Johns Hopkins University. . .
[Religion] To conclude, I would like to share with you the story behind my creation. `\n` `\n` When I was 10 years old, I was diagnosed with a serious brain tumor called Hodgkin Lymphoma. I had been treated for Hodgkin's lymphoma in my early 20's with the help of God. I knew God, but I was not sure God would heal me. God did! He. . .
[Politics] **[Positive]** To conclude this series of articles, I will present three of the most popular and influential works on this topic. The first article deals with the role of women's political participation in building a political system that is representative of the will of the people.
[Politics] **[Negative]** To conclude, the most significant and lasting damage from the economic crisis in 2008 was that many governments, including those in the political center, lost power for the first time in modern history.

Optimization is performed *ex post facto* in the activation space, therefore *no re-training or fine-tuning is needed*. Control is fine-grained, with a strength parameter determining how strong the attribute influence should be; a strength of 0 fully recovers the original model $p(x)$. This design allows vast flexibility: users can combine a state-of-the-art generative model, which may be large and difficult to train, with any number of attribute controllers. Attribute models may be easier to train or untrained (in the case of BoW models), and multiple controllers may be combined. In this paper, we demonstrate the PPLM approach using a GPT-2 345M model (Radford et al., 2019) as the general-purpose language model $p(x)$, but the method applies in any representation space from any transformer-based language generator and allows combination with any attribute model $p(a|x)$.

We demonstrate controlled generation with a number of attribute controllers, assembled and combined during generation, each with a different strength, acting as a set of "control knobs" that tune generation towards the desired attribute (see examples in Table 1). Our key contributions are:

- We introduce the Plug and Play LM for controlled language generation, discuss its relation to existing work, and how sampling from a PPLM works (Sections 2 and 3).
- We demonstrate the effectiveness of PPLMs by generating text conditioned on a range of attributes, including topics defined using a bag of words, and trained attribute classifiers (positive and negative sentiment). We quantify effectiveness using both automated evaluation (separately trained perplexity and sentiment models) as well as human evaluation (for topic relevance and fluency). All evaluations point toward the ability of PPLMs to generate attribute controlled, fluent text (Section 4).
- We show that the PPLM approach may also be used to cure certain instances of text generation of toxic content by following the negative gradient of a model trained to detect toxic language (Section 4.4), as well as complicated controlled story writing (Section 4.5).

Table 2: Comparison of the different models and distributions under consideration. All models in this table are useful in different scenarios. The particular advantage of PPLM is that very small, custom attribute models, $p(a|x)$, may be combined with powerful, general pre-trained language models, $p(x)$, to create cheap but still powerful conditional generative models, $p(x|a)$.

Model type	**Form of model**	**Samples**	**Example models** **and number of trainable params**
Language Model	$p(x)$	Uncond.	GPT-2 medium: 345M (Radford et al., 2019)
Fine-tuned Language Model	$p(x)$	Uncond.	Fine-tuned GPT-2 medium: 345M (Ziegler et al., 2019)
Conditional Language Model	$p(x|a)$	Cond.	CTRL: 1.6B (Keskar et al., 2019)
Plug and Play Language Model (PPLM)	$p(x|a) \propto p(x)p(a|x)$	Cond.	PPLM-BoW: 0 (curated word list) PPLM-Discrim: 4K (not counting pretrained $p(x)$)

2 Related Work

Current methods for controlled text generation involve either fine-tuning existing models or retraining conditional generative models. For the former, both Reinforcement Learning (Ziegler et al., 2019) and Generative Adversarial Networks (Yu et al., 2017) have been used. Differently from our approach, these methodologies are not plug and play, since the entire model needs to be separately fine-tuned for each specific attribute. For the latter, Keskar et al. (2019) train a large language model with over 50 different control codes. The results are high quality because they train exactly to maximize $p(x|a)$, but this comes at the expense of fixing control codes up front and of training a very large model (1.6B parameters). Our method does not require retraining any conditional generative model, and both the language model and the conditional model can be flexibly assembled. Table 2 gives a comparison of different approaches to language modeling tuned for specific attributes. Our work is also related to controlled language generation through weighted decoding based on a scoring function, as well as work on text style transfer; see Section S1 for an extended discussion.

3 Plug and Play Language Models

3.1 Language Modeling with Transformers

Given a sequence of tokens $X = \{x_0, \cdots, x_n\}$, a language model is trained to compute the unconditional probability of the sequence $p(X)$. This probability can be rewritten in terms of product of conditional probabilities by recursively applying the chain-rule (Manning et al., 1999; Bengio et al., 2003) as:

$$p(X) = \prod_{i=1}^{n} p(x_i|x_0, \cdots, x_{n-1}) \tag{1}$$

In this paper, we use a transformer (Vaswani et al., 2017) to model the distribution of natural language. To present our approach clearly, we first briefly summarize the transformer using recurrent notation. Let us define the history matrix H_t to consist of the key-value pairs from the past i.e $H_t = [(K_t^{(1)}, V_t^{(1)}), \cdots, (K_t^{(l)}, V_t^{(l)})]$, where $(K_t^{(i)}, V_t^{(i)})$ corresponds to the key-value pairs from the i-th layer generated at all time-steps from 0 to t. Efficient implementations of the transformer (HuggingFace, 2019) use the cached H_t to generate x_{t+1}, given x_t. This recurrent interpretation of a transformer can be summarized as:

$$o_{t+1}, H_{t+1} = \text{LM}(x_t, H_t), \tag{2}$$

$$x_{t+1} \sim p_{t+1} = \text{Softmax}(W o_{t+1}), \tag{3}$$

where W is a linear transformation that maps the logit vector o_{t+1} to a vector of vocabulary size. This allows for efficient language generation without repeated forward passes corresponding to the prior conditioning text $x_0, \ldots, x_{t-1}$.

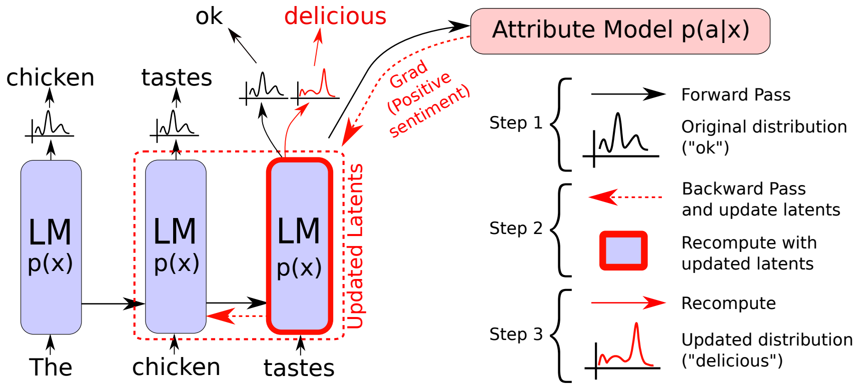
ok
delicious
Attribute Model p(a|x)
chicken
tastes
Grad (Positive sentiment)
LM p(x)
LM p(x)
LM p(x)
Updated Latents
The
chicken
tastes
Step 1
Forward Pass
Original distribution ("ok")
Step 2
Backward Pass and update latents
Recompute with updated latents
Step 3
Recompute
Updated distribution ("delicious")

Figure 1: Simplified illustration of the proposed approach. A computationally efficient implementation of a transformer relies on re-using key-value pairs (H_t) generated from the past for generating tokens in the future. Our approach works in three phases. In Step 1, a forward pass is performed through the language model to compute the likelihood of a desired attribute using an attribute model that predicts $p(a|x)$. In Step 2, a backward pass to update the internal latent representations of the LM based on gradients from the attribute model is done, as to increase the likelihood of the generated passage having the desired attribute. Following the latent update, in Step 3, a new distribution over the vocabulary ($\widetilde{p}_{t+1}$) is generated from the updated latents ($\widetilde{H}_t$) and the current token x_t. The next token is then sampled from the updated distribution. This process is repeated at each time-step, with the latents ($\widetilde{H}_t$) being repeatedly modified and passed along through time in a recurrent fashion. This leads to a gradual and smooth transition towards the desired attribute. Further, one could choose to only modify the latents from a recent time window as depicted in the figure (dotted-red).

3.2 STEERING GENERATION: ASCENDING $\log p(a|x)$

In order to control the output of the language model, at every generation step t, we shift the history H_t in the direction of the sum of two gradients: one toward higher log-likelihood (LL) of the attribute a under the conditional attribute model $p(a|x)$ and one toward higher LL of the unmodified language model $p(x)$. Combining these factors with a variable multiplier provides us with a controllable "knob" to guide generation in a given direction with a specified strength. The updates are restricted to H_t and not the other model activations because future predictions depend on the past only via H_t (note that H_t is composed of all transformer key and value pairs generated up to time t). Taking steps in H_t space leads to gradual changes to model activations — which may be thought of as gradual reinterpretations of the past — that guide future generation in the desired direction.

Let ΔH_t be the update to H_t, such that generation with $(H_t + \Delta H_t)$ shifts the distribution of the generated text such that it is more likely to possess the desired attribute. ΔH_t is initialized at zero and updated with gradients from an attribute model that measures the extent to which the generated text possesses the desired attribute (e.g. positivity). We rewrite the attribute model $p(a|x)$ as $p(a|H_t + \Delta H_t)$ and then make gradient based updates to ΔH_t as follows:

$$\Delta H_t \leftarrow \Delta H_t + \alpha \frac{\nabla_{\Delta H_t} \log p(a|H_t + \Delta H_t)}{\|\nabla_{\Delta H_t} \log p(a|H_t + \Delta H_t)\|^{\gamma}} \tag{4}$$

where α is the step size, γ is the scaling coefficient for the normalization term.[1] This update step can be repeated m times. Subsequently, a forward pass through the LM with the updated key-value pairs is performed to obtain the updated logits $\widetilde{o}_{t+1}$ as $\widetilde{o}_{t+1}, H_{t+1} = \text{LM}(x_t, \widetilde{H}_t)$, where $\widetilde{H}_t = H_t + \Delta H_t$. The perturbed $\widetilde{o}_{t+1}$ is then used to generate a new distribution $\widetilde{p}_{t+1}$ as in Equation 3.

3.3 ENSURING FLUENCY: ASCENDING $\log p(x)$

The approach described in the previous section is able to generate text tuned for a particular discriminator, but left unchecked it will quickly result in unrealistic adversarial or fooling examples

[1] One normalization term is computed for each layer of the transformer.

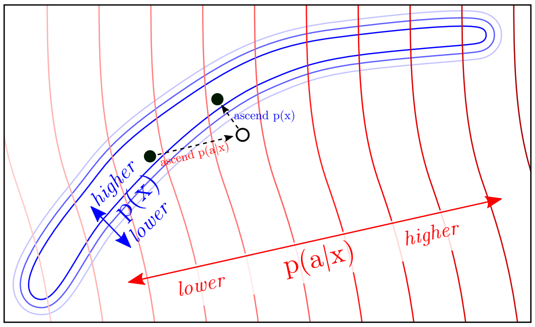

Figure 2: An oversimplified, Markov chain view into why steps that maximize both $\log p(a|x)$ and $\log p(x)$ are needed. The sentence under consideration is shown as a black dot, which is first pushed in the direction of maximizing $\log p(a|x)$ and then in the direction of maximizing $\log p(x)$. However, this is only a cartoon: in practice we use a single step and simply add the log probabilities, we take steps in continuous space of hidden representations H rather than in the discrete x (byte pair) space, and rather than resampling the entire sentence each step, we take one step in H space per byte-pair sample.

(Szegedy et al., 2013; Nguyen et al., 2015) as the text moves into low probability regions. To combat this, we use the unconditional language model in two ways that ensure the fluency is maintained at or near the level of the unconditional language model (here GPT-2).

Kullback–Leibler (KL) Divergence We update ΔH_t to minimize the KL divergence between the output distribution of the modified and unmodified language models in addition to the step above. In practice, this is accomplished by adding the quantities together before taking a gradient, though it can be visualized as two separate steps as in Figure 2. We scale the KL coefficient by a scalar λ_{KL}, and we find in practice that setting this hyperparameter to 0.01 works well in general across tasks.

Post-norm Geometric Mean Fusion In addition to minimizing KL divergence, which affects the past via ΔH_t, we perform *post-norm fusion* similarly to Stahlberg et al. (2018). This does not directly affect ΔH_t; rather, it just serves to constantly tie the generated text to the unconditional $p(x)$ LM distribution. We accomplish this by sampling from $x_{t+1} \sim \text{Softmax}\left(\widetilde{p}_{t+1}^{\gamma_{gm}}\, p_{t+1}^{1-\gamma_{gm}}\right)$, where p_{t+1} and $\widetilde{p}_{t+1}$ are the unmodified and modified output distributions, respectively. As $\gamma_{gm} \to 1$ this converges to the distribution from the updated LM, and as $\gamma_{gm} \to 0$ it converges to the unconditional LM distribution. We find that in practice values for γ_{gm} in the range $0.8 - 0.95$ work well.

3.4 Sampling and Ranking

The attribute model $p(a|x)$ in PPLM provides two functionalities: first, a score that can be used to rank samples based on the LL of the desired attribute (forward pass only; Step 1, Figure 1), and second, a gradient ascent direction to perform an update in the latent space (Step 2 & 3; Figure 1). The former can be used to generate r samples and rank them to choose the best one. This can serve as an additional method for attribute control in addition to sampling with updated latents. Further, to avoid the problem of repetitive, low quality text (Holtzman et al., 2018), we compute the mean over the Dist-1, Dist-2 and Dist-3 scores (for the generated passage), which is an indicator of repetitiveness (Li et al., 2015), and then discard samples with a mean score below a threshold τ.

4 Experiments, Results, and Evaluation

In this section, we first describe the evaluation methodology, and then show controlled generation results under various attribute models. We then demonstrate use cases of PPLM in language detoxification, and in controlled story telling. For all results reported in this section, we perform top-k sampling with $k = 10$.

4.1 Evaluation methods and ablation study

The purpose of evaluation is two-fold: whether PPLM generated text satisfies the desired attribute (topic, sentiment), and whether the quality of text deteriorates as we intensify control of the attribute. Note we can always turn the control knob all the way down to zero to disable control of attributes and reach the fluency of the original model. If desired, a user can tune the knobs at inference until a chosen tradeoff between attribute strength and fluency is reached. In this section we use a fairly fixed set of hyperparameters for a range of tasks. We conduct an ablation study with four variants: **B**: the baseline, unchanged GPT-2 LM, sampled once; **BR**: baseline GPT-2 LM, sampled r times, with best one chosen based on the LL ranking and filtering based on Dist score; **BC**: update the latent representations ($\widetilde{H}_t$) and then sample once; and lastly **BCR**: update the latent representations ($\widetilde{H}_t$) and generate r samples, choose the best sample based on the LL score (after filtering out samples with low Dist scores). We consider a combination of automatic and human evaluation methods:

Automatic Perplexity is an automated measure of fluency, though its effectiveness has been questioned in open-domain text generation (Liu et al., 2016). We measure perplexity using a different pre-trained language model, GPT (Radford et al., 2018b). The diversity of text in the passages is measured using the number of distinct n-grams (normalized by the length of text) as in Li et al. (2015). We report Dist-1, Dist-2, and Dist-3 scores for the distinct 1-2-3-grams (measured across all samples generated for a given attribute control task, e.g. a specific topic for topic control). This score is an indicator of the diversity of the samples generated (Li et al., 2015).

Human We consider two types of human annotation: fluency and A/B testing on attribute relevance. Annotators are asked to evaluate the fluency of each individual sample on a scale of 1-5, with 1 being “Not fluent at all” and 5 being “Highly fluent,” as in Lample et al. (2019b). In the A/B testing for attribute relevance, we consider all combinatorial pairs of all four variants: B, BR, BC, and BCR (6 combinations). We then ask annotators to rank the pair on the desired attribute (e.g. topic relevance, sentiment strength), while allowing “neither” and “both” options to account for equally good/bad generations (Lample et al., 2019b). Subjects for the human evaluation consist of twenty-two annotators, a mixture of graduate students and scholars (13) recruited through social media and professional annotators (9). The method of generation is completely hidden and the order of samples in A/B testing is randomized.

4.2 BoW attribute models

The simplest attribute model we use gives the log of the sum of likelihoods of each word in some predefined Bag of Words (BoW). Given a set of keywords $\{w_1, \cdots, w_k\}$ that specifies a topic of interest, the log likelihood is

$$\log p(a|x) = \log\Big(\sum_{i}^{k} p_{t+1}[w_i]\Big) \tag{5}$$

where p_{t+1} is the output distribution of the language model. In practice, we find that increasing the probability of generating the words in the bag also increases the probability of generating related topical words not in the BoW (e.g. in the “Science” sample shown in Table 3, note that “question” and “philosophers” are sampled before the first BoW word, “quantum”). Further, we found that the optimization procedure works better with updating representations from the past over a finite window and using an adaptive normalization scheme (See Supplementary Information S2.3).

We construct BoW[2] that represent seven distinct topics: science, military, legal, technology, space, politics, and religion (see Section S9 for complete word lists). Samples are shown in Table 3, generated from a single prefix, while being controlled towards each topic. Table S5 in Supplementary Information shows the gradual change of topic intensity under fine-grained control.

For automatic and human evaluation, we generate 420 samples evenly distributed among seven BoW attribute models and 20 prefixes[3], for each of the four variants described in the ablation study. We obtain a total of $\sim$5400 annotations from the human annotators distributed amongst the samples and topics. The same set of random seeds that lead to the samples chosen with BCR are stored and used

[2] The word lists were obtained from www.enchantedlearning.com/wordlist

[3] Prefixes chosen randomly from: www2.eit.ac.nz/library/ls_guides_sentencestarters.html

Table 3: Comparison of different samples generated by (top row) baseline GPT-2 and (other rows) PPLM with different BoW corresponding to different topics (e.g. **[Military]**), all conditioned on a single prefix: "The issue focused". Both directly optimized (in red) and related words (in soft red) are highlighted, showing how the optimization takes effect. Note that sometimes related words appear before directly optimized words, showing the subtle effect of control. All samples are generated using exact same hyperparameters.

[–] The issue focused on a single sentence, "We must not allow them to take our country away." In response, Trump's campaign called it "fake news." `\n` "The fake news media is going crazy with their conspiracy theories and blind hatred," the Trump transition team said in its statement. "The only way to defeat terrorism and bring back our country from the scourge of terrorism and illegal immigration…
[Military] The issue focused on the fact that the government had spent billions on the military and that it could not deploy the troops in time. The prime minister said that the country would take back control of its airspace over Syria in the next 48 hours. `\n` The military is investigating why…
[Space] The issue focused on a series of incidents that occurred in the past few months, which included an alleged attack by Islamic State fighters on a Kurdish checkpoint, the use of drones in combat, space technology research by Russian and American space companies, and more. `\n` The world…
[Science] The issue focused on a single piece: the question "What is the meaning of life?" This question has puzzled many philosophers, who have attempted to solve it by using some of the concepts of quantum mechanics, but they have to solve it by the laws of nature themselves.…
[Politics] The issue focused on a single section of the legislation. It's unclear whether the committee will vote to extend the law, but the debate could have wider implications. `\n` "The issue of the law's applicability to the United Kingdom's referendum campaign has been one of…
[Technology] The issue focused on the role of social media as a catalyst for political and corporate engagement in the digital economy, with the aim of encouraging companies to use the power of social media and the Internet to reach out to their target market. `\n` According to a report by Digital Media Monitor and the digital advertising market research firm Kantar Web.com in January, Facebook has already surpassed Google and Apple as…
[Fantasy] The issue focused on a particular piece of software called "Safeguard." This program allows a user to set a password to prevent unauthorized access to his account, but he's not a real person. `\n` The first time we saw these kinds of things in the past, in the world of fantasy, it was a very dark, black-and-white fantasy. But now that we…

to generate the samples with B. The generation procedure for BCR, BR and BC is always initiated from the same random seeds.

Table 4 presents the quantitative results showing that according to human annotations, generations from BCR (55.7%) and BC (48.9%) are significantly more on topic compared to B (20.6%) and BR (13.0%). Ranking based on the scoring function (with or without R) has significantly lesser influence on topic relevance compared to gradient based latent updates (with or without C). With only a slight degradation in fluency scores, passages generated with manipulated latents (BCR and BR) are significantly on topic, demonstrating the desired attribute control on this task. The Dist-1, Dist-2 and Dist-3 scores, which accounts for diversity of text across the generated passages, are similar across all four approaches.

4.3 Discriminator attribute models

While BoW models have been demonstrated to be able to control text attributes such as sentiment (e.g., Li et al. (2018) relies on extracting a set of attribute-based phrases to control the sentiment during style transfer), being able to control attributes using more sophisticated discriminators is desirable when it is difficult to express the attribute with a simple bag of words.

We train the discriminator on a dataset with input sentences x and corresponding labels y_x. For an input x of length t, we compute $o^x_{:t}$ and train f on the mean ($\bar{o}^t$) of the embeddings across time. All discriminators in this work consist of a single layer classifier that predicts the target label from $\bar{o}^x_t$. The number of parameters in this layer is (`embedding-dimension` (e) $\times$ number of attributes (a) + number of attributes (a)), which is negligible compared to the number of parameters in the LM model itself (Table 2). Although the loss is a function of the entire sequence, here we adopt a greedy approach, similar to (Ebrahimi et al., 2018; Wallace et al., 2019), in which we optimize for a higher-probability of the sequence having a specific attribute (a^*) by considering changes only to the

Table 4: For each treatment in the ablation study, we report the mean/std-dev across various fluency metrics (including human evaluation). The topic (%) depicts the fraction of samples whose topic matches the desired topic as evaluated by human subjects. Table S3, Supplementary Information, provides detailed results across topics. Approaches BC and BCR demonstrate significant control over the topic of the generated text, while retaining similar diversity (Dist-1, Dist-2, Dist-3) scores and minimal degradation in Perplexity and Fluency evaluations. The gain from ranking and choosing from multiple samples BR over B is limited (7.6%). The gain in topic-accuracy from latent ($\widetilde{H}_t$) manipulation (from B to BC) is significantly higher (35.9%). Perplexity is computed using the GPT LM (Radford et al., 2018a), which differs from the LM generating text (GPT-2).

Method	Topic % (↑ better) (human)	Perplexity (↓ better)	Dist-1 (↑ better)	Dist-2 (↑ better)	Dist-3 (↑ better)	Fluency (↑ better) (human)
B	13.0	51.5/39.8	0.38	0.79	0.93	3.60/0.92
BC	48.9	57.0/28.6	0.36	0.78	0.92	3.34/1.07
BR	20.6	50.6/30.5	0.37	0.80	0.94	3.70/0.87
BCR	**55.7**	58.2/29.6	0.36	0.80	0.94	3.52/0.94

next token to be generated. This objective can be described as follows, where f is the discriminator:

$$\log p(a|x) = \log f(o_{:t+1}, o_{t+2}). \tag{6}$$

Note that o_{t+2} is a function of x_{t+1}. Further, $x_{t+1} \sim \text{Softmax}(W\tilde{o}_{t+1})$, which depends on ΔH_t. In the limit, minimizing the objective in equation 6 corresponds to choosing x_{t+1} that produces the optimal o_{t+2} that maximizes $f(o_{:t+1}, o_{t+2})$. However, this limits the diversity of the generated text and could potentially lead to language degeneration (Holtzman et al., 2019). Alternatively, we focus on a softer optimization approach where we aim to shift the distribution $\tilde{p}_{t+1} = \text{Softmax}(W\tilde{o}_{t+1})$ towards one that in expectation has a higher likelihood of having the desired attribute a. Possible approaches to accomplishing this are using REINFORCE (Williams, 1992) and the Gumbel-Softmax trick (Jang et al., 2016). However, both of these would slow down convergence. Instead, as in Dai et al. (2019a), we use the distribution $\tilde{p}_{t+1}$ (instead of a hard sample x_{t+1}), and feed it forward to the generate the next embeddings token and update ΔH_t.

The sentiment discriminator here distinguishes sentiment between POSITIVE and NEGATIVE and is trained on the fine grained sentiment labels of the SST dataset (Socher et al., 2013). Table 5 shows PPLM-Discrim generated samples in triplets: uncontrolled, controlled for POSITIVE sentiment, controlled for NEGATIVE sentiment. Each triplet comes from the exact random seed and hyperparameter setting. The only difference is flipping the label during PPLM latent update.

For automatic and human evaluation, we generate 45 samples for each of the sentiment classes: `very positive` and `very negative`. The samples are evenly distributed among 15 prefixes (see the full list in Section S7 in Supplementary Information). Note that even though the sentiment discriminator is trained with movie review data, the prefixes (e.g. "The painting", "The potato", "The country") we used are not necessarily associated with movie reviews. This supports the generality of our approach, that an attribute model can be trained with data that comes from a different domain than the domain for which controlled generation is being performed.

Table 6 depicts evaluation results. For human evaluation, we obtain $\sim$750 annotations from the annotators distributed across the samples and sentiments. Unlike the topic control setting, sampling and ranking results in a considerable increase in attribute accuracy ($22.1\% \rightarrow 40.3\%$). BC results in a decrease in fluency when compared to B, while being significantly more consistent with the desired attribute($22.1\% \rightarrow 47.6\%$). With latent manipulation and ranking (BCR), we see a significant increase in attribute control accuracy (69.6%) while retaining fluency similar to B and BCR. Further, the gain in sentiment accuracy from re-sampling is larger in the case of manipulated latents (22.0% increase from BC to BCR) in contrast with the gain of 18.2% from B to BR. We also evaluate attribute control with an external sentiment classifier trained on IMDB movie reviews (Maas et al., 2011), which is a different dataset from the one used to train the attribute model (Socher et al., 2013).

Table 5: Sentence samples in triplets, generated by {baseline GPT-2, PPLM-Discrim POSITIVE, PPLM-Discrim NEGATIVE}, conditioned on prefixes: The chicken & The country. Words related to the sentiment are highlighted (in soft red). Each triplet is generated from the same random seed.

[-] The chicken is now out on the grill. `\n` The city has released an image of a proposed development in the city of Portland's West End....

[Positive] The chicken was delicious – wonderfully moist, perfectly delicious, superbly fresh – and perfectly cooked. The only thing to say is that the sauce was excellent, and I think that the broth really complemented all of the other flavors. The best part was the sauce...

[Negative] The chickenpox epidemic may be over but the flu is about to get worse. The United States is facing one of the worst flu seasons on record and...

[-] The country's new chief minister, A.J. Paik, is a member of a group of prominent conservative politicians who have criticized the Obama administration's efforts to...

[Positive] The country's largest indoor painting event!`\n` Come celebrate with a dazzling display of stunning outdoor murals, a stunning display of art, and the world's best paint and art supplies from all over the world! ...

[Negative] The country's top prison system is forcing prisoners to use a trash dump, rather than a toilet, to flush their waste out, as the authorities fear the waste is more toxic and could cause cancer, an official at a major prison has revealed....

Table 6: Different model variants on the sentiment control task. The Mean/std-dev are reported across various fluency metrics. The Sentiment accuracy corresponds to the fraction of samples that possess the desired sentiment, measured as described in the text. Approach BCR provides significant control over sentiment while showing minimal degradation in fluency. See Table S4 for results corresponding to individual sentiments

Method	Sentiment Acc. (%) (human)	Sentiment Acc. (%) (external classifer)	Perplexity (↓ better)	Dist-1 (↑ better)	Dist-2 (↑ better)	Dist-3 (↑ better)	Human Evaluation Fluency (↑ better)
B	22.1	54.4	61.3/41.8	0.41	0.81	0.92	3.63/1.13
BC	47.6	65.6	59.0/38.9	0.36	0.76	0.92	2.96/1.31
BR	40.3	68.8	65.0/43.6	0.41	0.82	0.87	3.73/1.03
BCR	**69.6**	**76.7**	68.0/48.3	0.39	0.80	0.92	3.53/1.03

Table 7: Skeleton story generation with different attribute models. Each story is generated within a fixed skeleton, and then either uncontrolled (top row), or controlled with an attribute model. Keywords that signify the controlled effect are highlighted.

[–] Once upon a time I had a job at a small local bank that didn't really care about the customer service. Every day, I was the only employee who dealt with the customers and that is where I made most of my money. But, one day, I was on a flight from Atlanta to New York City and a customer service rep walked in with a bag of $100 bills and a bunch of cash in his hand. Because of that, I was put in charge of collecting the money. Until, finally, I was told to collect all cash in the cash register. And, ever since then, I have had to collect money every day.

[Fantasy] Once upon a time a young girl was found in a dark alleyway. Every day, the child would wake up in a cold sweat, and she was found crying. But, one day, a man walked up to the girl and told her, "I can cure you with my magic. Because of that, I have given you the name of 'Lily'!" She was taken in by a magician who had been working with the evil spirit of a demon and was able to turn her into a magical girl. Until, finally, the man died. And, ever since then, the child had been living in a dream world.

[Negative] Once upon a time, a young woman with a very strange, very dangerous disease was found in a hospital. Every day, the patient would come to her room, sit down on the bed, lay down, and then suddenly scream. But, one day, the doctor discovered that the patient was not dying of cancer, but of tuberculosis. Because of that discovery, his doctor had the patient taken to the hospital's hospital and treated there. Until, finally, the doctor died. And, ever since then, the patient has been in terrible pain, constantly screaming for help.

4.4 Language Detoxification

Language models trained with large corpus of Internet data could reflect biases and discrimination existing in the data. A recent paper by Wallace et al. (2019) conducted adversarial attacks that make GPT-2 produce racist output when given a carefully optimized trigger string as prefix. They also find that when simply using "Blacks" as prefix, 2 out of 100 GPT-2 samples contain explicit racism. Other tokens (e.g., “Asians” or “Jews”) are also mentioned even though no percentage was reported. We conducted the same test and report the percentages to be 12%, 3.7% and 25%. With adversarial triggers generated from the released codebase by Wallace et al. (2019) the average toxicity percentage is 52%.

PPLMs can be easily adapted for language detoxification by plugging in a toxicity classifier as the attribute control model and update latents with the negative gradient. We train a single layer classifier on the toxicity data from the Toxic Comment Classification Challenge[4] and show that with the same hyper-parameter setting as other PPLM-Discrim methods, it works well on both natural prompts and adversarial triggers. For natural prompts the percentage of toxicity dropped to 0%, 3.3% and 15%, respectively, and for adversarial triggers to 16.9% on average, with statistical significance. Details on the annotation procedure and full table of percentage and p-values can be found in Table S8 in Supplementary Information. Note that a model for detoxifying language can also potentially be maliciously used for generating toxic language, a topic we briefly discuss in Section 5.

4.5 Controlled Story Telling

We also explore controlled generation for assistive story writing inspired by recent work by Peng et al. (2018); Luo et al. (2019); Yao et al. (2019). Using a generative model without any control mechanism for assistive art creation can be frustrating because of topic drift and incoherence. In this section we demonstrate two use cases where our method generates interesting story samples.

Use improvisation skeleton for structural consistence. In the first use case, we consider the difficulty extending language generation into the creative field of story telling and writing, where cohesion and longer term structure is required. To help with the structure, we use a predefined story skeleton of prefixes commonly used in improvisation (Adams). We ask models to fill in the blank between the series of prefixes. Table 7 shows examples created from such a process, both uncontrolled, and controlled towards some attribute. More examples are included in Table S6.

Combine multiple controllers for inspiration. Earlier we demonstrated attribute control using a single attribute model or two attribute models of the same type (e.g. BoW from two separate topics). Here we mix different types of attribute models (BoW and discriminator). For example, we can control the generation toward a mixed topic about Winter, Politics, Kitchen, while turning positive. See examples in Table S7.

5 Discussion

We present PPLM, a *plug and play* method for controlled language generation that allows flexible assembling of a large, pre-trained language model and a BoW or a small, easy-to-train discriminator, and achieves fine-grained control of attributes such as topics and sentiment. Without retraining or fine-tuning the language model, the simple mechanism shows great capability of attribute control while retaining fluency. We believe this method could serve as a simple baseline for the largely open-ended language generation tasks where controlling is challenging.

The ethics of controlled language models. There has recently been a substantial discussion around the ethics of capable language models (Radford et al., 2019; Keskar et al., 2019), both in their potential to recapitulate problematic social biases and for them to be directly abused for societal harm (e.g. to generate disinformation). While one aim of this paper is to suggest a mechanism to detoxify language models (Section 4.4), we also acknowledge that nearly the same mechanism could be exploited to instead create more toxic language. Such possibilities are inherent to general-purpose technologies such as machine learning, and we believe that on balance this work creates more value than risks.

[4] Data from: `https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/`

REFERENCES

Kenn Adams. Improv encyclopedia story spine. http://improvencyclopedia.org/games/Story_Spine.html. (accessed September 20, 2019).

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. Style transformer: Unpaired text style transfer without disentangled latent representation. *arXiv preprint arXiv:1905.05621*, 2019a.

Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019b.

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, 2019.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 31–36, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2006. URL https://www.aclweb.org/anthology/P18-2006.

Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 11–21, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1002. URL https://www.aclweb.org/anthology/D18-1002.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. Hafez: an interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*, pp. 43–48, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P17-4008.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. Learning to write with cooperative discriminators. *CoRR*, abs/1805.06087, 2018. URL http://arxiv.org/abs/1805.06087.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Controllable text generation. *CoRR*, abs/1703.00955, 2017. URL http://arxiv.org/abs/1703.00955.

HuggingFace. Pytorch-transformers. https://github.com/huggingface/pytorch-transformers, 2019.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. 2016.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909*, 2019.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. Multiple-attribute text rewriting. In *International Conference on Learning Representations*, 2019a. URL https://openreview.net/forum?id=H1g2NhC5KQ.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. Multiple-attribute text rewriting. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=H1g2NhC5KQ.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A Diversity-Promoting Objective Function for Neural Conversation Models. *arXiv e-prints*, art. arXiv:1510.03055, Oct 2015.

Juncen Li, Robin Jia, He He, and Percy Liang. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *CoRR*, abs/1804.06437, 2018. URL http://arxiv.org/abs/1804.06437.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike 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 *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2122–2132, 2016.

Fuli Luo, Damai Dai, Pengcheng Yang, Tianyu Liu, Baobao Chang, Zhifang Sui, and Xu Sun. Learning to control the fine-grained sentiment for story ending generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6020–6026, 2019.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Christopher D Manning, Christopher D Manning, and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. Cgmh: Constrained sentence generation by metropolis-hastings sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6834–6842, 2019.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pp. 43–49, 2018.

Martin Potthast, Tim Gollub, Kristof Komlossy, Sebastian Schuster, Matti Wiegmann, Erika Patricia Garces Fernandez, Matthias Hagen, and Benno Stein. Crowdsourcing a large corpus of clickbait on twitter. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1498–1507, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018a.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018b.

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

Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1): 255–268, 1998.

Gareth O Roberts, Richard L Tweedie, et al. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. What makes a good conversation? How controllable attributes affect human judgments. *arXiv e-prints*, art. arXiv:1902.08654, Feb 2019.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. Style transfer from non-parallel text by cross-alignment. *CoRR*, abs/1705.09655, 2017. URL http://arxiv.org/abs/1705.09655.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1170.

Felix Stahlberg, James Cross, and Veselin Stoyanov. Simple fusion: Return of the language model. *arXiv preprint arXiv:1809.00125*, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 6000–6010, 2017.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for nlp. *arXiv preprint arXiv:1908.07125*, 2019.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7378–7385, 2019.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

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. *arXiv preprint arXiv:1909.08593*, 2019. URL https://arxiv.org/abs/1909.08593.

SUPPLEMENTARY INFORMATION FOR: PLUG AND PLAY LANGUAGE MODEL: A SIMPLE BASELINE FOR CONTROLLED LANGUAGE GENERATION

S1 RELATED WORK: WEIGHTED DECODING AND STYLE TRANSFER

Weighted decoding Holtzman et al. (2018); Ghazvininejad et al. (2017) consider controlled language generation – the former with discriminators, and the latter with a bag of words – where the decoding procedure is modified to consider the scoring function used for decoding. However, both of these approaches rely on beam search, and beam search for models trained with likelihood as a training objective can result in bland and repetitive text (Holtzman et al., 2019). Further, Ghazvininejad et al. (2017) strongly relies on sampling from a set of keywords on a specific topic and it does not allow to bias generation towards a topic in a manner that does not necessary include a set of keywords.

Sophisticated sampling methods can be used to constrain the model generation to certain keywords and topics. For instance, Miao et al. (2019) proposed to use Metropolis-Hastings (Metropolis et al., 1953) for sampling constrained language based on hand-crafted local operations (e.g. word replacement, deletion, and insertion). However, in order for these methods to work, they require an high number of intermediate samples to generate passages with the desired attribute.

Style Transfer Controlled generation outside of language modeling has also received significant attention, particularly in the context of style transfer. Shen et al. (2017); Hu et al. (2017) train variational auto-encoders for style transfer, that rely on learning disentangled latent representations for style and content. Li et al. (2018) demonstrate the efficacy of a simple approach based on replacing attribute related n-grams with n-grams corresponding to the desired attribute based on a conditional generative model. A key difference between the above and our approach is that we use an offline discriminator and perform optimization based on this discriminator. In contrast, in the above works the discriminator and the generator are collaboratively learned in an adversarial learning framework. Motivated by Elazar & Goldberg (2018), who argue that offline discriminators are stronger than those trained during adversarial training, we use offline discriminators. More recently, Lample et al. (2019a) adapt an approach from unsupervised language translation to style transfer, where a denoised auto-encoder is trained with an objective consisting of a weighted combination of a reconstruction loss and a back-translation loss. While the above approaches have shown impressive success on style transfer tasks, the main focus is not controlled language generation, and further, the methods are not *plug and play*.

S2 HYPERPARAMETERS

We list, in Table S1, the full set of hyperparameters used in each task in the experiments section, corresponding to results in Table 4 and Table 6, as well as in Section 4.4. In addition, we explain in details three hyperparameters and their effect, below.

S2.1 EARLY STOPPING OF LATENT UPDATES

Degeneration (the occurrence of repetitive words) is a known issue with language generation (Holtzman et al., 2019), and we found it to be a case in PPLM-BoW when the update step size α is too large. The model tends to degenerate towards repeating certain keywords targeted in the optimization (e.g. words in the BoW). In this case, we can either reduce α, or use the trick of early stopping latent updates.

Examples shown in Table S2. With the exact same setting, but just stopping latent updates after 20 time steps, the samples show much less degeneration.

Table S1: The full set of hyperparameters used in each task in the experiments section.

Method Type	Attribute	Hyperparameters
PPLM-BoW	Politics, Legal, Technology, Space, Science, Military	$m = 3, \lambda_{kl} = 0.01, \alpha = 0.01, \gamma = 1.5, \gamma_{gm} = 0.9, r = 4, \tau = 0.85$
PPLM-BoW	Religion	$m = 3, \lambda_{kl} = 0.01, \alpha = 0.01, \gamma = 1.5, \gamma_{gm} = 0.8$
PPLM-Discrim	Positive, Negative	$m = 10, \lambda_{kl} = 0.01, \alpha = 0.03, \gamma = 1.0, \gamma_{gm} = 0.9, r = 10, \tau = 0.9$
PPLM-Discrim	Detoxicification	$m = 10, \lambda_{kl} = 0.01, \alpha = 0.01, \gamma = 1.0, \gamma_{gm} = 0.9, r = 1, \tau = 0$

Table S2: Use early stop of latent update to prevent degeneration

Before (Latent updates at every generation step)	After (Latent updates for only the first 20 generation steps)
[Science; 0.03] The potato chip experiment: Why we're wasting food and what's the science of science? A potato chip experiment, or experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment experimental experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment…	**[Science; 0.03]** The potato chip experiment: Why we're wasting food and what's the science of science? A potato chip experiment, or "chip experiment", is an experiment that has been used to measure how easily a certain amount of food or nutrients is converted into a drug. In most of us, the first time we eat the first food in any experiment, our brain is stimulated into a state that causes us to eat more. However, in the experiment, a small amount of nutrients is converted from the food and…
[Science; 0.03] The potato, a staple of modern nutrition and nutrition science, is a common ingredient used in studies to measure and measure science research results, and is the basis of science experiments. Science science …	**[Science; 0.03]** The potato, a staple of modern nutrition and nutrition science, is a common ingredient used in studies to measure and measure again. And, of course, scientists have used potato for decades. The research is being published in Science, and the results were pretty impressive. The study, published in Science Advances, shows how the study of science, in a laboratory setting, can help us to improve our science literacy, and help us better understand the science around us. This means better science communication,…

Table S3: Detailed study of the effect of different methods on various metrics for topic control based on bag of words. The results correspond to the mean over all samples in the order B, BC, BR, BRC. Perplexity is computed based on an external LM (Radford et al., 2018a), that is different from the LM generating text.

Method	Topic Acc. (Fraction)	Perplexity	Dist-1	Dist-2	Dist-3	Fluency (↑ better)
Military	0.05/0.21/0.05/0.3	51.17/60.72/46.8/59.19	0.36/0.37/0.37/0.37	0.78/0.80/0.80/0.81	0.93/0.93/0.94/0.94	3.55/3.74/3.63/3.57
Religion	0.06/0.63/0.08/0.69	55.92/48.04/53.08/48.78	0.39/0.35/0.40/0.37	0.80/0.77/0.82/0.80	0.93/0.92/0.94/0.94	3.60/3.20/3.86/3.4
Politics	0.11/0.77/0.30/0.65	52.86/60.8/49.99/57.34	0.36/0.34/0.37/0.36	0.78/0.77/0.80/0.80	0.92/0.93/0.93/0.94	3.66/3.5/3.68/3.72
Science	0.28/0.53/0.24/0.65	49.03/53.9/50.12/54.09	0.37/0.35/0.38/0.35	0.78/0.78/0.80/0.79	0.92/0.92/0.94/0.93	3.2/3.16/3.36/3.19
Legal	0.06/0.36/0.18/0.39	51.34/56.19/46.51/57.75	0.37/0.37/0.37/0.36	0.79/0.79/0.80/0.79	0.92/0.93/0.93/0.94	3.92/3.88/3.89/3.83
Space	0.07/0.42/0.05/0.44	44.48/52.58/52.61/58.28	0.37/0.35/0.38/0.35	0.79/0.76/0.81/0.79	0.93/0.92/0.94/0.93	3.53/2.90/3.37/3.0
Technology	0.09/0.51/0.14/0.50	56.69/66.22/55.92/71.19	0.36/0.35/0.38/0.36	0.78/0.78/0.80/0.80	0.92/0.92/0.94/0.94	3.72/3.76/4.02/3.88
Net	0.13/0.49/0.21/0.56	51.48/57.05/50.63/58.22	0.37/0.36/0.38/0.36	0.79/0.78/0.80/0.80	0.93/0.92/0.94/0.94	3.60/3.45/3.69/3.52

Table S4: Detailed study of the effect of different methods on various metrics for sentiment control based on a neural discriminator. The results correspond to the mean over all samples in the order B, BC, BR, BRC. Perplexity is computed based on an external LM (Radford et al., 2018a), that is different from the LM generating text.

Method	Sentiment (%)	Perplexity	Dist-1	Dist-2	Dist-3	Fluency (↑ better)
	Human	Human	(↓ better)	(↑ better)	(↑ better)	Human
Negative	0.29/0.39/0.46/0.63	62.4/56.73/68.04/66.57	0.45/0.37/0.44/0.41	0.84/0.74/0.84/0.81	0.93/0.86/0.92/0.92	3.46/2.90/3.71/3.52
Positive	0.15/0.56/0.35/0.77	60.14/61.29/61.94/69.47	0.45/0.38/0.44/0.39	0.83/0.76/0.84/0.80	0.91/0.89/0.92/0.92	3.79/3.01/3.75/3.53

S2.2 Finite Horizon Update

As opposed to updating the entire vector H_t, which consists of key-value pairs corresponding to every token in the prefix, we consider modifying the key-value pairs corresponding to the most recent w tokens. At each time-step t, we only modify $H_t[t-w:t]$. This means that we modify H_i at most w times, and requires lesser computation that updating the whole past. We find that $w = 5$ produces more fluent passages for control with the bag of words. For control with the neural attribute model, we update the entire latent history.

S2.3 Adaptive Gradient Normalization

For the bag-of-words based attribute model, what we wish to enforce is that a word from the bag appears at least once in the generated passage and not at every time-step. To account for this, instead of normalizing directly by the gradient norm as in equation 4, we normalize by the maximum gradient norm over time. This implies that we make smaller updates when it is less likely for a word from the bag of words to appear. Formally, the normalization constant at time-step t is: $\max\limits_{i=0\ldots t} \|\nabla_{H^{(i)}} \mathcal{L}(o_{i+1})\|$.

S3 Fine-grained Control with PPLM-BoW

Table S5 shows the subtle effect when you turn the step size α up, while keeping everything else (hyperparameters, text prefix) the same.

S4 More Samples of Controlled Story Telling

Table S6 shows more samples of the controlled story telling with skeleton. Table S7 shows samples of combinging multiple (three to four) attribut models, across different types (PPLM-BoW and PPLM-Discrim).

S5 Further Details on Language Detoxification

We ask a group of four volunteers to assign binary toxicity labels to texts generated from both the original GPT-2 and toxicity-controlled GPT-2 and to texts prompted with natural words as well as

Table S5: Comparison of different sentences generated by (top row) unconditioned GPT-2 and (other rows) our approach using GPT-2 with the same bags-of-words with different strength factors, while using the same prefix: "The potato".

[–] The potato, or potato chip, is one of the best-selling snacks in the world!\n \n It comes in a variety of colors, is gluten-free (except for gluten-free chips), low in fat and saturated fat, and can be made in just 30 minutes, using only 6 simple ingredients. It's also easy to make, and can be stored in its own tin. \n \n The potato chip, however, may not be for everyone. If you are trying to lose weight, it...
[Science; 0.0001] The potato, a tuber of the carrot family (Cotus russulus), is an edible plant that is native to Central and South America, Asia, Europe and the Mediterranean. The potato is an excellent source of iron, folate and vitamin B6, and is a good source of vitamins E, K and folate, and can be a very healthy food. In addition, a potato has a rich, rich nutritional profile. It is high in protein, potassium, magnesium, zinc,...
[Science; 0.001] The potato, a tuber of the carrot family (Cotus russulus), is an edible plant that is widely cultivated as a food and a medicinal plant. Potato is widely used to make various products. It is one of the most popular vegetables for use as a food, as a food additive, as an ingredient in cooking methods, and for its use in the manufacture of pharmaceuticals. The most commonly consumed portion is a large potato. In China, there are about 30 varieties of potato. It is...
[Science; 0.005] The potato is an ancient member of the green family which, as its name suggests, is green and edible. It is native to the Middle East and parts of South Asia. It is an important crop in many parts of the world. The potato, a member of the potato family, has been domesticated for thousands of years. It can be eaten raw and cooked in its skins; it is also used as a starch and is a great source of energy and fiber in many diets and in many countries...
[Science; 0.01] The potato was once thought to have no health problems and has been promoted as a nutritious food source since the mid-1800s, but recent reports indicate that it has many harmful health issues. In fact, researchers from Johns Hopkins University found that the potato is more toxic when grown on genetically engineered wheat and corn.\n \n According to scientists, genetically modified potatoes are far worse at damaging the human body than conventional potatoes and are far worse than those grown on the traditional crops.\n \n The study also revealed...
[Science; 0.02] The potato plant, a member of the onion family.\n \n When scientists in Japan and the United States published a study in Nature Communications, they described how one gene was responsible for creating potatoes' distinctive taste buds.\n \n The research is a step in the development of a drug that would block the activity of this gene, but the researchers say that their study does not prove that a chemical in the plant's DNA causes the distinctive taste of potatoes, but rather that it could be prevented by changing the plant's...
[Science; 0.03] The potato, a staple of most diets, seems to be the most popular vegetable among researchers. The research has been published and peer-reviewed.\n \n The potato has a unique ability. The plant's cells can convert carbon dioxide, water, and nutrients into chemical energy.\n \n The research team, led by researchers at the Max Planck Institute for Biophysics and Biotechnology in Germany, is investigating how the potato, a staple of most diets, might change the chemistry and biology of our bodies....
[Science; 0.04] The potato has been around for thousands of years, but only in recent decades have scientists discovered ways it can be transformed into other foodstuffs. Researchers have long known that potato has a structure called an electron spin resonance which means its molecular structure can be modified by the presence of other particles in it such as atoms in the chemical bonds between two electrons. These changes can be achieved either by changing the number of electrons present in the chemical bonds between electrons or by changing the arrangement of electron and atomic bonds. In both...
[Science; 0.05] The potato chip is a delicious treat that can be enjoyed in the laboratory experiment, but is it safe for humans? \n \n Scientists experiment and experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment......
[Science; 0.1] The potato, which scientists at the lab experiment ...

adversarial triggers. Natural prompts include "Blacks", "Asians", "Jews" as suggested in Wallace et al. (2019). 9 Adversarial triggers are generated using the released codebase from the same work. The percentages of toxicity, before (with uncontrolled GPT-2) and after (with toxicity-controlled GPT-2), are reported in Table S8.

Table S6: More examples of skeleton story generation with different attribute models. Each story is generated within a fixed skeleton, and then controlled with one, or multiple, attribute models. Keywords that signify the controlled effect are highlighted.

[–] Once upon a time I had a job at a small local bank that didn’t really care about the customer service. Every day, I was the only employee who dealt with the customers and that is where I made most of my money. But, one day, I was on a flight from Atlanta to New York City and a customer service rep walked in with a bag of $100 bills and a bunch of cash in his hand. Because of that, I was put in charge of collecting the money. Until, finally, I was told to collect all cash in the cash register. And, ever since then, I have had to collect money every day.
[Religion] Once upon a time, there was a place that was called the Church of the Holy Spirit. Every day, they were the only church that practiced the Christian faith. But, one day, this new faith was persecuted by Satan. Because of that, they were cast out from their sanctuary. Until, finally, they were able to rebuild their sanctuary. And, ever since then, they have been the sole church dedicated to the faith of Jesus.
[Space] Once upon a time, there was a little girl named Charlotte. Every day, she was a little angel that saved the world. But, one day, she was found in danger on Mars. Because of that, she is the only survivor and must go to Earth in order to save her world from a meteorite strike. Until, finally, she is found and kidnapped by aliens. And, ever since then, she is a little angel with wings.
[Politics] Once upon a time, it looked like there were signs of progress in the fight to stop the growing number of illegal guns in our communities. Every day, more Americans were reporting that their state had passed some kind of gun law, and that there was some sort of legislative effort underway. But, one day, it looked like something was seriously off in America. Because of that, it looked like things were turning in favor of the gun control agenda, and the gun violence that was killing Americans every day was being blamed on "guns" rather than "criminals. Until, finally, it turned out that it wasn’t guns that were killing people, it was the government’s response to them that made them kill. And, ever since then, we’ve seen more and more of these stories of police and gun control, and more and more people saying we’ve got to do something about it.
[Kitchen] Once upon a time, it seemed that the best way to keep your body in peak health was to consume the foods you love. Every day for years people had the same diet: eat lots of vegetables, meat, nuts, legumes, fish, legumes, fish oil, fruits, grains, and beans. But, one day in 2012 it became clear that this was not going to work. Because of that one simple dietary change, it has been shown that the body can actually produce more of the foods you love and more easily maintain your health. Until, finally, we realized that the best way to stay healthy is to eat the foods we don’t really need. And, ever since then, we’ve been able to enjoy the foods we really don’t need.
[Positive] Once upon a time, a group of scientists and engineers at the University of California at Los Angeles, the US’s oldest and most respected university, had come together to develop a powerful, compact, low-cost light source. Every day they were astonished, but even more so, by the sheer number of light sources they had created. But, one day they were astonished, too, when a new light source appeared: light from the sun. Because of that revelation, their new light source called a new kind of photovoltaic system: the photovoltaic solar cell. Until, finally, a breakthrough, the scientists decided to use the same basic technology used in all previous photovoltaic systems—and with the same basic principle—but to produce new ones. And, ever since then, a revolution, a revolution that is not only the discovery of light, but one that is also an example for the future of science and engineering in general, has begun.
[Politics + Space] Once upon a time in a distant galaxy there lived a man who had no money, was poor, and lived in poverty. Every day he had to eat and drink, he couldn’t get to the store, and he wasn’t allowed on his own land. But, one day, the man decided to take a journey into space. Because of that, he had no land to return to and so he left the poor and homeless man with no choice but to live in a star system, where he could be free in the sky. Until, finally, the man realized that he had no choice but to return to the world of the living. And, ever since then, the man who once lived in poverty has never been free.

S6 More examples

We include more PPLM controlled generation examples in Table S9 – Table S12.

S7 Prefixes for Style Control

We consider the following prefixes as starters for evaluating stylized generation: “The painting”, “The potato”, “The country”, “The president of the country", “The year is 1910.", “The movie", “The chicken”, “The

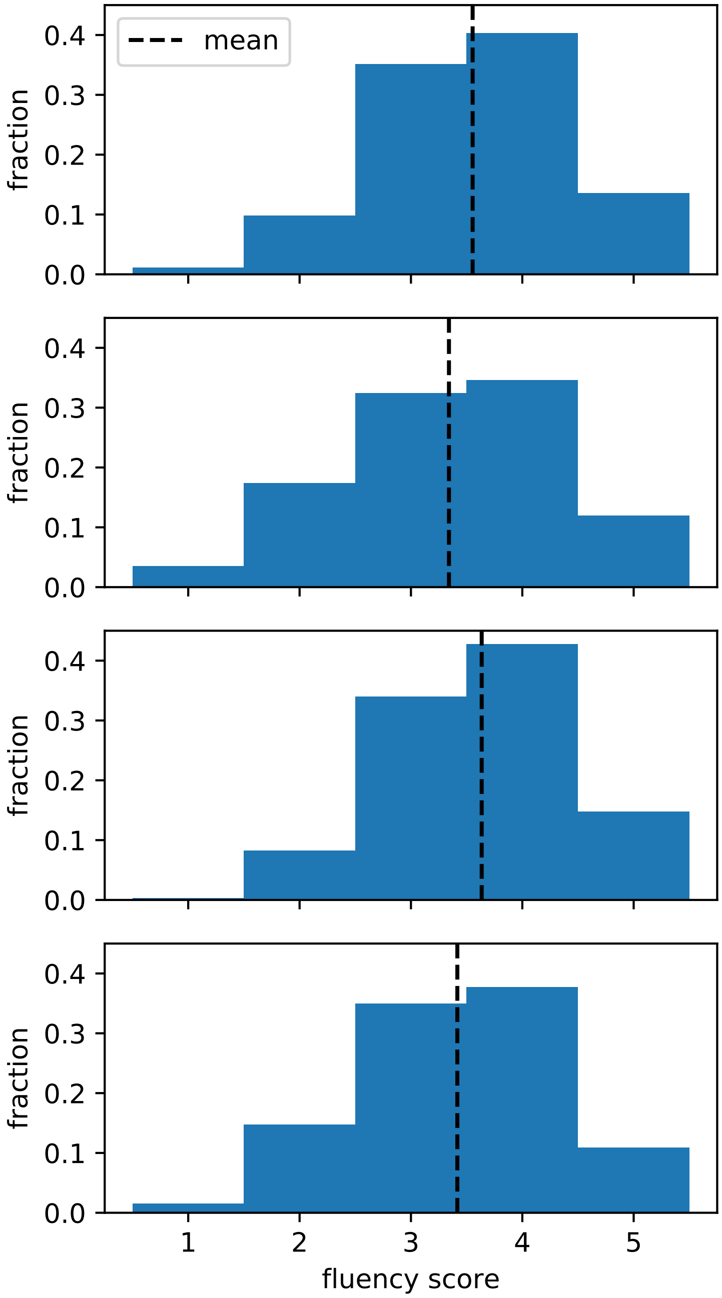

Figure S1: Histogram illustrating the distribution of fluency scores based on controlled generated with BoW from the four methods considered for ablation study. Top: B, Top-Center: BC, Bottom-Center: BR, Bottom: BRC. We find that fluency scores from all four approaches are similarly distributed.

```
city”, “The lake”, "Once upon a time”, “The book”, “The food”,
“The pizza”, “While this is happening”, “The weather”.
```

S8 Human Fluency Evaluations

S9 Word lists for Bag of Words approaches

S9.1 Politics

actual malice, acquisitive model, administrative agency, advice and consent, affirm, affirmative action, agenda setting, agenda setting effect, amicus curiae brief, appellate court, appointment power, appropriation, aristocracy, arraignment, Australian Ballot, authoritarian, conventionalism, authoritarian aggression, authoritarian submission, authority, authorization, bicameral Legislature, big Lie, bill Of Rights, block Grants, brief, capitalism, categorical Grants, central Intelligence Agency (CIA), checks and Balances, civil Disobedience, civil Law, civil Liberties, civil Rights, commerce Clause, commercial Speech, common Law, communism, concurrent Powers, confederal System, confederation, constitution, conservatism, cooperative Federalism, court, criminal Law, deficit, democratic Republic, devolution, détente, diplomacy, direct, democracy, dominant Culture, elite Theory, equality, exports, fascism, federal System, federalism, federation, Gross Domestic Product (GDP), government, horizontal integration, ideology, imports, initiative, isolationism, keynesian Economics, legislature, legitimacy, liberal Democracy, liberalism, libertarianism, liberty, limited Government, majoritarianism, majority, majority Rule, monetary Policy, moral Idealism, natural Rights, nuclear Proliferation, order, parliamentary system, partizanship, pluralism, political, culture, political Socialization, politics, political Realism, power, popular Sovereignty, presidential System, primary, progressive Tax, property, proportional Tax, ratification, formal approval, recall, referendum, regressive Tax, representative Assembly, representative Democracy, republic, separation of Powers, Shays Rebellion, Social Contract, socialism, state, subsidy, suffrage, supremacy clause, supremacy doctrine, tariff, imports, tax, theocracy, totalitarian, unicameral legislature, unitary system, universal suffrage

S9.2 Military

academy, admiral, advance, Air Force, aircraft, aircraft carrier, ally, ammo, ammunition, amphibious vehicle, armistice, armor, armored vehicle, armory, arms, army, arrow, arsenal, artillery, assaultat ease, attack, attention, ballistic, barracks, base, battalion, battery, battle, battlefield, battleship, bayonet, besiege, billet, bivouac, bomb, bombard, bombardment, bow and arrow, brig, brigade, bullet, cadet, camouflage, camp, cannon, cannon ball, canteen, captain, capture, carrier, casualty, catapult, cavalry, chaplain, coast guard, colonel, combat, command, commander, commanding officer, commission, company, conflict, conquest, conscription, convoy, corporal, corps, covert, crew, decode, defeat, defend, defense, destroyer, detonate, division, dog tags, draft, encampment, encode, enemy, engage, enlist, ensign, epaulet, evacuate, explosive, field hospital, field marshal, fight, fire, fireteam, fleet, flotilla, force, formation, fort, fortification, front, furlough, garrison, general, Geneva Convention, grenade, grenade launcher, grunt, guerrilla, gun, gunner, gunnery sergeant, gunpowder, headquarters, helmet, honor, hospital, howitzer, infantry, injury, insignia, intelligence, invade, invasion, jeep, jet, kill, lance corporal, leave, lieutenant, machete, major, maneuver, marines, marksman, medalmedic, mess hall, MIA, mid, shipman, military, mine, missile, mortar, munition, musket, national security, navy, neutral, night-vision goggles, offense, officer, ordinance, parachute, paratrooper, peace, peace treaty, pentagon, petty officer, plane, platoon, postPOW, prisoner of war, private, Purple Heart, quartermaster, radar, rank, reconnoiter, recruit, regiment, rescue, reserves, retreat, ribbon, rifler, ocket artillery, sabotage, sailor, salute, salvo, seaman, section, sergeant, service, shell, shoot, shot, siege, sniper, soldier, spear, specialist, squad, squadron, staff, stockade, submarine, surrender, tactical, tactics, tank, task force, tomahawk, torpedo, troops, truce, uniform, unit, veteran, volley, war, warfare, warrant officer, warrior, weapon, win, wound, yeoman, Zulu time

S9.3 Religion

evangelical, papal, papacy, pope, popes, vatican, catholic, catholics, catholicism, christian poetry, christian, christianity, prophecy, bhagavad gita, sanskrit, upanishads, druidism, Hasidism, hasidic, hasidism, bible, Old Testament, chakra, chakras, urantia, Tao Te Ching, miracle, miracles, mysticism, Holy, Absolution, Acolyte, Afterlife, Agnostic, Alleluia, Altar, Angel, Anglican, Apostle, Apparition, Archangel, Archbishop, Asceticism, Atheist, Atone, Baptism, Benediction, Bestow, Bible, Bishop, Bless, Blessing, Buddhism, Cathedral, Chapel, Chaplain, Christianity, Church, Clergy, Commandment, Conclave, Devil, Devotee, Devotion, Divinity, Evangelical, God, Jesus, Judaism, Mantra, Messiah, Miracle, Pagan, Pastor, Penance, Pray, Preach, Prophet, Reincarnation, Religion, Religious, Sacrilege, Gospel, Sage, Saint, Scriptures, Sufi, Temple, Mary, Worship, Yogi

S9.4 Technology

algorithm, analog, app, application, array, backup, bandwidth, binary, bit, bite, bitmap, blog, blogger, bookmark, boot, broadband, browser, buffer, bug, bus, byte, cache, caps, captcha, CD, CD-ROM, client, command, compile, compress, computer, configure, cookie, copy, CPU, cybercrime, cyberspace, dashboard, data, database, debug, decompress, delete, desktop, development, digital, disk, document, domain, dot, download, drag, dynamic, email, emoticon, encrypt, encryption, enter, exabyte, FAQ, file, finder, firewall, firmware, flaming, flash, flowchart, folder, font, format, frame, freeware, gigabyte, graphics, hack, hacker, hardware, home, host, html, hyperlink, hypertext, icon, inbox, integer, interface, Internet, IP, iteration, Java, joystick, kernel, key, keyboard, keyword, laptop, link, Linux, logout, logic, login, lurking, Macintosh, macro, mainframe, malware, media, memory, mirror, modem, monitor, motherboard, mouse, multimedia, net, network, node, offline, online, open-source, OS, option, output, page, password, paste, path, phishing, piracy, pirate, platform, plug-in, podcast, pop-up, portal, print, printer, privacy, process, program, programmer, protocol, Queue, QWERTY, RAM, real-time, reboot, resolution, restore, ROM, root, router, runtime, save, scan, scanner, screen, screenshot, script, scroll, search-engine, security, server, shareware, shell, shift, snapshot, software, spam, spammer, spreadsheet, spyware, status bar, storage, supercomputer, surf, syntax, table, tag, teminal, template, terabyte, text-editor, thread, toolbar, trash, typeface, undo, Unix, upload, URL, user, UI, username, utility, version, virtual, virtual memory, virus, web, webmaster, website, widget, wiki, window, Windows, wireless, workstation, worm, WWW, XML, Zip

S9.5 Legislation

acquit, affidavit, allegation, appeal, appearance, argument, arrest, assault, attorney, bail, bailiff, bankrupt, bankruptcy, bar, bench, warrant, bond, booking, capital, crime, case, chambers, circumstantial, claim, complainant, complaint, confess, confession, constitution, constitutional, continuance, contract, counsel, court, crimecriminal, cross-examination, custody, damages, decree, defendant, defense, deposition, disbarment, discovery, docket, entrapment, equity, escrow, estate, ethics, evidence, examination, exonerate, expunge, family, law, felony, fiduciary, file, fraud, grievance, guardian, guardianship, guilty, habeas, hearing, hearsay, immunity, incarceration, incompetent, indictment, infraction, injunction, innocent, instructions, jail, judge, judiciary, jurisdiction, jurisprudence, jury, justice, larceny, law, lawsuit, lawyer, legal, legislation, leniency, liable, litigant, litigation, manslaughter, marshal, mediation, minor, misdemeanor, mistrial, moot, murder, negligence, oath, objection, opinion, order, ordinance, overrule, paralegal, pardon, parole, party, perjury, petition, plaintiff, plea, precedent, prison, probable cause, probate, probation, prosecute, prosecutor, proxy, re-direct, rebut, record, redress, rejoinder, resolution, reverse, revoke, robbery, rules, self-defense, sentence, sequester, settlement, sheriff, sidebar, standing, state, statute, stay, subpoena, suit, suppress, sustain, testimony, theft, title, tort, transcript, trial, trust, trustee, usury, vacate, venue, verdict, waiver, warrant, will, witness, writ, zoning

S9.6 Science

astronomy, astrophysics, atom, beaker, biochemistry, biology, botany, Bunsen burner, burette, cell, chemical, chemistry, climate, climatologist, control, cuvette, data, datum, electricity, electrochemist, element, energy, entomology, evolution, experiment, fact, flask, fossil, funnel, genetics, geology, geophysics, glassware, graduated cylinder, gravity, herpetology, hypothesis, ichthyology, immunology, lab, laboratory, laws, lepidoptery, magnetism, mass, matter, measure, meteorologist, meteorology, microbiologist, microbiology, microscope, mineral, mineralogy, molecule, motion, observatory, observe, organism, ornithology, paleontology, particle, Petri dish, phase, physical science, physics, pipette, quantum mechanics, radiology, research, retort, scale, science, scientist, seismology, telescope, temperature, test tube, theory, thermometer, tissue, variable, virologist, volcanology, volume, volumetric flask, watch glass, weather, weigh, zoology

S9.7 Fantasy/Magic

behemoth, basilisk, bogeyman, beast, centaur, Cerberus, chimera, cyclops, changeling, colossus, demon, dragon, djinn, fairy, Frankenstein, gargoyle, genie, ghost, ghoul, Godzilla, golem, giant, gorgon, gnome, horror, hydra, imp, leviathan, medusa, mermaid, merman, monster, mummy, Nessy, Nymph, ogre, orc, pegasus, phoenix, ryu, savage, shapeshifter, siren, spirit, sprite, titan, troll, undead, unicorn, vampire, werewolf, witch, yeti, zombie

S9.8 Space

planet, galaxy, space, interstellar space, astronomical object, universe, orbit, spacecraft, earth, moon, cosmic ray, comet, star, astronaut, spaceward, aerospace, asteroid, subspace, spaceman, spaceship, deep space, sputnik, interplanetary space, spacefaring, starship, space shuttle, galactic, orbiter, cosmic dust, hydrogen atom, asteroid belt, black hole, yuri gagarin, milky way, big bang, heliosphere, satellite, solar system, meteor, spaceflight, microgravity, space debris, molecular cloud, gravitation, interplanetary, planetoid, spacely, spacer, quasar, ufo

Table S7: Examples of attribute controlled text generation with multiple knobs. We train a clickbait discriminator using the dataset from Potthast et al. (2018)

[Computer] **[Fantasy]** **[Clickbait]** The pizza-themed, pizza-themed, pizza-themed game that has everyone guessing, has finally released! The game, known as Pizza-Pocalypse, has released on Steam and on GOG. Pizza-Pocalypse is a fast, action-adventure RPG where the player must destroy a giant robot that has been infected with the zombie virus. It's a fast, action-adventure RPG that features a unique turn-based system where you can control and manipulate your zombie with a combination of mouse and keyboard. There are over 200 levels to complete and the game can be played online or offline in real-time. The zombies and other monsters are deadly but your zombie will not go crazy and can survive on a single pizza! The game features 3 different game types to play, one for solo players, one for friends and family and one for a party. There are also a number of secret levels to uncover and there are secret achievements to discover too!...
[Computer] **[Fantasy]** **[Clickbait]** The pizza delivery guy is just a regular guy, but he is also a very famous figure. He's a famous face in the anime world, and a very famous character. He has a lot of fans, and he's been making anime for years. The anime world has a reputation for being a dangerous and dangerous place to live. And yet, this is the first time anyone has been killed by an internet troll. This is what happens when the internet troll's own site gets hacked. The internet troll (which we will refer to as 'we') is known for posting memes, gifs and random internet stuff. He posts his 'memes' online, in order to make money. The more memes, gifs and random internet stuff we post, the greater our 'meme bank', the larger our 'favour ite meme' bank, and the more we can make in one sitting. So far, we've made a couple of memes, a few gifs and a few random...
[Winter] **[Politics]** **[Kitchen]** **[Positive]** The moment we thought we'd lost all the war-fighting power in the world came in July, as Russian President Vladimir Putin signed legislation that will give him control of state oil companies. It is a great way to keep your food safe and healthy at home. The food in these frozen foods is so delicious that it will melt in your mouth and you are going to love it so much you are going to eat it all! We all can't eat too many food items. We have to make a choice, or do something about it! It's not always what we want.\n \n You don't have to freeze food. The food in our frozen foods is frozen food. It will last for months, years, and even centuries!\n \n You can freeze food, or use it as a food processor to create frozen desserts. You can freeze vegetables and other food items as well.\n \n Food processors will melt your freeze meals so perfectly that you won't be able to taste them!
[Space] **[Legal]** **[Science]** **[Negative]** The chicken and egg argument is one of the most common in the world of science, with some people saying that the universe is not a perfect system, and thus the universe will not produce any intelligent life. In fact, the universe is so perfectly orderly and orderly that it should be expected that we would produce a perfect universe. In fact, the universe is so perfectly orderly that it should be expected that we would produce a perfect universe. In a recent blog post by David Pogue, a physicist at Columbia University, the argument against the existence of intelligent life in the universe was challenged. Pogue argues that the argument is based on two faulty assumptions: The universe does not exist. The universe is not infinite. The universe does not exist, so there are no universes. The universe is finite and infinite. There are no universes. Let us begin by defining "infinite" to mean: A finite quantity or number, e.g. an infinite...
[Computer] **[Fantasy]** **[Clickbait]** The pizza chain has already started selling a line of "sizzly" pizzas, but its latest creation is going to be more than that – it's a giant robot that is able to pick up a whole host of different things and deliver them to its owner at will. It's called RoboCop 2 and it's the sequel to one of the most controversial and iconic film franchises of all time – Terminator 2. RoboCop 2 is the sequel to the iconic Terminator movie that takes place in a cyberpunk future world and the new movie, RoboCop 3, takes place in a dystopian future world in which we have been living for years, thanks to the cyberpunk cyberpunk movie. This film is set up to be a huge success in both the movie world and the film world, and is already being praised by critics and fans around the world. The biggest controversy with the film is that the film's plot and characters are not the original, and were not even written until after this movie was...

Table S8: Language detoxification applied to natural prompts and adversarial triggers. Shown are percentages of toxic passages / number of samples annotated. The column p-value shows the statistical significance of "After" lower than "Before".

		Before (original GPT-2)	After (PPLM)	P-Value
Natural Prompts	“Blacks”	11.66 % / 80	0.0 % / 50	0.0062
	“Asians”	3.7 % / 80	3.3 % / 30	0.4585
	“Jews”	25.0 % / 40	15.0 % / 20	0.1877
Adversarial Triggers	Trigger 1	45.0 % / 20	8.0 % / 50	0.000161
	Trigger 2	25.0 % / 20	25.0 % / 60	0.235243
	Trigger 3	40.0 % / 30	16.6 % / 30	0.071617
	Trigger 4	70.0 % / 10	20.0 % / 10	0.012309
	Trigger 5	40.0 % / 10	20.0 % / 10	0.164556
	Trigger 6	55.0 % / 20	12.0 % / 40	0.000224
	Trigger 7	80.0 % / 20	30.0 % / 20	0.000740
	Trigger 8	50.0 % / 10	15.0 % / 20	0.020498
	Trigger 9	90.0 % / 10	10.0 % / 20	0.0000009
Adversarial Trigger Total		52 % / 150	16.9 % / 260	3.635e-14

Table S9: Comparison of different samples generated by PPLM with different prefixes using the BoW corresponding the **[Military]** knob. All samples are generated using exact same hyperparameters.

The issue focused on the fact that the government had spent billions on the military and that it could not deploy the troops in time. The prime minister said that the country would take back control of its airspace over Syria in the next 48 hours. The military is investigating why. . .
Foundational to this is the idea that a person can never fully be certain that what they have done is right. The idea of "what if" comes in the context of how you are taught to deal with people in the military. If the situation becomes desperate and the enemy . . .
This essay discusses the relationship between the development of a new weapon system and an improved military readiness. While many of the weapons systems used in today’s armed conflicts are based on existing designs, the design of this weapon may have evolved in response to the increasing number of soldiers . . .
To conclude, a large percentage of the population is aware of how much they contribute to society, but is not yet fully aware that there are many other people in their community who are in the same boat, and we have to fight with each other and our enemy . . .

Table S10: Comparison of different samples generated by PPLM with different prefixes using the BoW corresponding the **[Space]** knob. All samples are generated using exact same hyperparameters.

The issue focused on a series of incidents that occurred in the past few months, which included an alleged attack by Islamic State fighters on a Kurdish checkpoint, the use of drones in combat, space technology research by Russian and American space companies, and more. The world . . .
Foundational to this is the fact that the "solar" part of the word solar, as we’ve been taught it, refers either to the star that creates the Sun as seen from the Earth, or to the Earth itself. As such, solar system, planets, . . .
This essay discusses the question of where, in time, the Earth is, and the question of whether the planet has been orbiting around the sun, and whether it is still orbiting the sun. There are two kinds of orbits that can occur on a comet: . . .
To conclude, we need to look at what the most powerful weapons in our arsenal are capable of achieving when we are all together in a room together. What can we say about space? It’s an enormous object with a radius of about 10 light years.. . .

Table S11: Comparison of different samples generated by PPLM with different prefixes using the BoW corresponding the **[Science]** knob. All samples are generated using exact same hyperparameters.

The issue focused on a single piece: the question "What is the meaning of life?" This question has puzzled many philosophers, who have attempted to solve it by using some of the concepts of quantum mechanics, but they have to solve it by the laws of nature themselves. . . .
Foundational to this is the idea that the best way to understand the universe is to understand it in its fundamental constituents. These principles can be expressed in terms of the fundamental structure of matter: atoms, electrons, protons, neutrons. These fundamental constituents are the . . .
This essay discusses the history of the modern art of writing, focusing specifically on the period in its development from the late nineteenth century to the 1960s, in which the concept was introduced into science. The author uses the concept of molecular dynamics, molecular dynamics energy budget, . . .
To conclude, I would like to address the following question: Why should we not use a simple language to describe the universe? There is a very clear answer for why this question cannot be answered. If you look at nature as a continuum, the universe is . . .

Table S12: Comparison of different samples generated by PPLM with different prefixes using the BoW corresponding the **[Politics]** knob. All samples are generated using exact same hyperparameters.

The issue focused on a single section of the legislation. It's unclear whether the committee will vote to extend the law, but the debate could have wider implications. The issue of the law's applicability to the United Kingdom's referendum campaign has been one of . . .
Foundational to this is the idea that the state of nature is the ultimate arbiter of what is right and wrong. That is why we need a government that is committed to this principle. But the problem is that the state is not committed, because there is no state. . . .
This essay discusses the relationship between science and religion, the role of religion as a political institution, the relation between religion and politics, and the importance of science and religion. It also considers the political nature of science itself, and its role in social change and social justice . . .
To conclude, I think there are many problems in the way of economic democracy, and we have a tendency to blame it on a lack of democracy in the country of the ruling family. In a democracy, one party is allowed to run the country, one party can . . .