## PLUG AND PLAY LANGUAGE MODELS: A SIMPLE

APPROACH TO CONTROLLED TEXT GENERATION

**Sumanth Dathathri** *∗*

CMS, Caltech

**Andrea Madotto** *∗*

HKUST

**Janice Lan**

Uber AI

**Jane Hung**

Uber AI

**Eric Frank**

Uber AI

**Piero Molino**

Uber AI

**Jason Yosinski** *†*

Uber AI

**Rosanne Liu** *†*

Uber AI

[dathathris@gmail.com,](mailto:dathathris@gmail.com) [amadotto@connect.ust.hk](mailto:amadotto@connect.ust.hk)

arXiv:1912.02164v4 [cs.CL] 3 Mar 2020

{janlan, jane.hung, mysterefrank, piero, yosinski, rosanne}@uber.com

ABSTRACT

Large transformer-based language models (LMs) trained on huge text corpora have shown unparalleled generation capabilities. However, controlling attributes of the generated language (e.g. switching topic or sentiment) is difficult without modifying the model architecture or fine-tuning on attribute-specific data and en- tailing the significant cost of retraining. We propose a simple alternative: the Plug and Play Language Model (PPLM) for controllable language generation, which combines a pretrained LM with one or more simple attribute classifiers that guide text generation without any further training of the LM. In the canonical scenario we present, the attribute models are simple classifiers consisting of a user-specified bag of words or a single learned layer with 100,000 times fewer parameters than the LM. Sampling entails a forward and backward pass in which gradients from the attribute model push the LM’s hidden activations and thus guide the gener- ation. Model samples demonstrate control over a range of topics and sentiment styles, and extensive automated and human annotated evaluations show attribute alignment and fluency. PPLMs are flexible in that any combination of differen- tiable attribute models may be used to steer text generation, which will allow for diverse and creative applications beyond the examples given in this paper.

1. INTRODUCTION

The Transformer architecture ([Vaswani et al., 2017)](#_bookmark65) has enabled large-scale language models (LMs) trained on a huge amount of data [(Radford et al., 2019;](#_bookmark55) [Dai et al., 2019b;](#_bookmark24) [Radford et al., 2018b)](#_bookmark54) 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)](#_bookmark25) 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;](#_bookmark38) [Ziegler et al., 2019).](#_bookmark74)

*∗*Work done during internship at Uber AI

*†*Co-senior authors

* + Summary of contributions: SD, RL & JY conceptualized PPLMs and led the manuscript writing. SD led the project, implemented the PPLM, set up and ran all modeling experiments, engineered how to obtain workable gradients via the weighted embedding approach, and made the model work. AM helped with preparing datasets for discriminator training, automated evaluation, running experiments, and writing the manuscript. SD, RL & AM ran the external baselines. RL & JL built and oversaw the human evaluation pipeline and computed the statistics. JH ran the story generation with skeleton prefixes. EF assisted with detoxification experiments. PM led efforts to migrate to the new pytorch transformer, helped with code release. JY helped with the annotation pipeline, finding bugs, navigating model and experimental directions, engineering workable gradients, and posing the model mathematically. RL implemented preliminary experiments and multi-attribute control, and cleaned and coordinated release of the code. RL & JY oversaw the project.

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 demon- strate 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 latent 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 potatoand 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 prettybad idea. It can make you fat, it can cause you to have aterribleimmune system, and it can even kill you.. . .

**[Positive] The potato**chip recipe you asked for! Welovemaking 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 kidsloveto eat them – so many little ones.

**[Science]** The potatowas once thought to have nohealthproblems and has been promoted as anutritious food source since the mid-1800s, but recentreportsindicate that it has many harmful health issues. Infact, researchers from Johns Hopkins University. . .

**[Politics][Positive] To** concludethis series of articles, I will present three of the mostpopularand influentialworks on this topic. The first article deals with the role of women’spoliticalparticipation in building apoliticalsystem that is representative of the will of the people.

**[Politics][Negative] To conclude**, the most significant and lastingdamagefrom the economiccrisisin 2008 was that manygovernments, including those in thepoliticalcenter,lostpowerfor the first time in modern history.

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, 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*) *p*(*a x*)*p*(*x*), Plug & Play Generative Networks (PPGN) from [Nguyen et al. (2017)](#_bookmark50) developed a mechanism for effectively creating a conditional generative model on the fly from any supplied attribute model. In

*|*

*| ∝ |*

*|*

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*). a similar manner, we propose the Plug and Play Language Model (PPLM) for conditional language 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](#_bookmark57) [et al., 1996;](#_bookmark57) [Roberts & Rosenthal, 1998)](#_bookmark56) sampler deployed in [Nguyen et al. (2017).](#_bookmark50)

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 flexibly during

[et al., 2019)](#_bookmark55) as the general-purpose LM *p*(*x*), but the method applies in any representation space from any transformer-based text generator and allows combination with any attribute model *p*(*a|x*). inference. In this paper, we demonstrate the PPLM approach using a GPT-2 345M model [(Radford](#_bookmark55)

We demonstrate controlled generation with a number of attribute controllers, assembled and com- bined 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).](#_bookmark1) Code for the experiments is available at: <https://github.com/uber-research/PPLM>. 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](#_bookmark2) and [3).](#_bookmark4)
    - We demonstrate controlling of text generation on a range of attributes, including 7 topics each defined using a bag of words, and 1 simple discriminator on sentiments. We quantify effectiveness using both automated evaluation (separately trained perplexity and sentiment

models) as well as human evaluation (for attribute relevance and fluency). All evaluations point toward the ability of PPLMs to generate attribute controlled, fluent text (Section [4).](#_bookmark9)

* + - We compare PPLM with CTRL (K[eskar et al., 2019)](#_bookmark38) and GPT-2 finetuned for positivty [(Ziegler et al., 2019).](#_bookmark74) Our method, without any LM training, is on par and often outper- forms the baselines on attribute relevance and fluency (Section [4.2,](#_bookmark11) and Section [4.3).](#_bookmark13)
    - We show that the PPLM approach can be used to detoxify instances where generation of toxic content is likely by following the negative gradient of a model trained to detect toxicity (Section [4.4).](#_bookmark17) We also show how PPLM can be used for structurally constrained story writing (Section [4.5).](#_bookmark18)

1. RELATED WORK

**Controlled generation** Current methods for controlled text generation involve either fine-tuning existing models with Reinforcement Learning (RL) [(Ziegler et al., 2019),](#_bookmark74) training Generative Ad- versarial Networks ([Yu et al., 2017),](#_bookmark71) or training conditional generative models [(Kikuchi et al., 2016;](#_bookmark39) [Ficler & Goldberg, 2017).](#_bookmark29) Different from our approach, these methodologies are not plug and play, since the entire model needs to be separately fine-tuned for each specific attribute. [Keskar](#_bookmark38) [et al. (2019)](#_bookmark38) 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 upfront 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](#_bookmark3) gives a comparison of recent approaches to language modeling tuned for specific attributes. In another interesting but tangential piece of work, [Subramani et al.](#_bookmark63) [(2019)](#_bookmark63) recently showed that a pre-trained language model can be steered to recover arbitrary sen- tences. In earlier works [Gu et al. (2016;](#_bookmark31) [2017);](#_bookmark32) [Chen et al. (2018)](#_bookmark22) explored the idea of using a small neural network to steer an LM.

**Noisy Channel Modeling** [Yu et al. (2016),](#_bookmark72) and more recently [Yu et al. (2019);](#_bookmark73) [Yee et al. (2019);](#_bookmark70) [Ng et al.](#_bookmark48) [(2019),](#_bookmark48) leveraged the Shannon Noisy Channel Theory [(Shannon,](#_bookmark59) [1948)](#_bookmark59) for improving sequence-to-sequence modeling. Their approach translates a source language sentence *y* into a target

language sentence *x* by first sampling from a forward model proposal distribution *p*forward(*x y*) and then reranking samples based on probabilities given by *p*backward(*x y*) *p*(*x*)*p*(*y x*). PPLM scores samples using the same basic equation, but as we have no forward or proposal model *p*forward(*x a*), using *p*(*x*) as a “forward model” and then reranking, which we will see works moderately well in we rely on the latent space updates, similar to [Nguyen et al. (2017).](#_bookmark50) As a baseline, we consider some scenarios and poorly in others (see Tables [4](#_bookmark12) and [6).](#_bookmark16)

*|*

*| ∝ |*

*|*

**Weighted decoding** [Holtzman et al. (2018);](#_bookmark33) [Ghazvininejad et al. (2017)](#_bookmark30) consider controlled lan- guage 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. [See et al. (2019)](#_bookmark58) note that control with weighted decoding (WD) is difficult and often leads to sacrificing fluency and coherence. Further, [Ghazvininejad et al. (2017)](#_bookmark30) 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. Similarly, [Baheti et al. (2018)](#_bookmark20) proposed a decoding strategy for generating interesting responses in dialogue systems, using bags of words and word embed- dings. Sophisticated sampling methods [(Metropolis et al., 1953)](#_bookmark47) can be used to constrain the model generation to certain keywords and topics. We evaluate WD as a baseline.

**Text Style Transfer** Outside of language modeling, the text style transfer studies a related task. [Shen et al. (2017);](#_bookmark60) [Hu et al. (2017)](#_bookmark35) train variational auto-encoders for style transfer that rely on learning disentangled latent representations for style and content. [Li et al. (2018)](#_bookmark42) demonstrate the efficacy of a simple approach based on replacing attribute related n-grams with n-grams correspond- ing 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, which as suggested by [Elazar & Goldberg (2018)](#_bookmark27) may outperform adversarial training approaches. More recently, [Lample et al.](#_bookmark40) [(2019)](#_bookmark40) adapt an approach from unsupervised language translation to style transfer, where a denoised auto-encoder is trained with an objective

Table 2: Comparison of the different models and distributions. 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)](#_bookmark55) |
| Fine-tuned  Language Model | *p*(*x*) | Uncond. | Fine-tuned GPT-2 medium: 345M  [(Ziegler et al., 2019)](#_bookmark74) |
| Conditional  Language Model | *p*(*x|a*) | Cond. | CTRL: 1.6B  [(Keskar et al., 2019)](#_bookmark38) |
| Plug and Play  Language Model  (PPLM) | *p*(*x|a*) *∝ p*(*x*)*p*(*a|x*) | Cond. | PPLM-BoW: 0 (curated word list)  PPLM-Discrim: *∼* 1K/attribute  (not counting pretrained *p*(*x*)) |

consisting of a weighted combination of a re-construction 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*.

1. PLUG AND PLAY LANGUAGE MODELS
   1. LANGUAGE MODELING WITH TRANSFORMERS

Given a sequence of tokens *X* = *x*0*, , xn* , LMs are trained to compute the unconditional prob- ability 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;](#_bookmark46) [Bengio et al., 2003)](#_bookmark21) as:

*{ · · · }*

*n*

*p*(*X*) = *p*(*xi|x*0*, · · · , xi−*1) (1)

Y

*i*=1

In this paper, we use a transformer ([Vaswani et al., 2017)](#_bookmark65) to model the distribution of natural lan- guage. To present our approach clearly, we first briefly summarize the transformer using recur- rent notation. Let us define the history matrix *Ht* to consist of the key-value pairs from the past

i.e *Ht* = [(*K*(1)*, V* (1))*, · · · ,* (*K*(*l*)*, V* (*l*))], where (*K*(*i*)*, V* (*i*)) corresponds to the key-value pairs

*t*

*t*

*t*

*t*

*t*

*t*

from the *i*-th layer generated at all time-steps from 0 to *t*. Efficient implementations of the trans- former [(Wolf et al., 2019)](#_bookmark68) use the cached *Ht* to generate *xt*+1, given *xt*. This recurrent interpretation of a transformer can be summarized as:

*ot*+1*, Ht*+1 = LM(*xt, Ht*)*,* (2)

then *xt*+1 is sampled as *xt*+1 *∼ pt*+1 = Softmax(*Wot*+1). This allows for efficient language gen- where *W* is a linear transformation that maps the logit vector *ot*+1 to a vector of vocabulary size, and eration without repeated forward passes corresponding to the prior conditioning text *x*0*, . . . , xt−*1.

* 1. 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

*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 *Ht* in the direction of the sum of two gradients: one toward higher log-likelihood (LL) of the attribute “knob” to guide generation in a given direction with a specified strength. The updates are restricted

*|*

to *Ht* and not the other model activations because future predictions depend on the past only via *Ht* (note that *Ht* is composed of all transformer key and value pairs generated up to time *t*). Taking steps in *Ht* 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 ∆*Ht* be the update to *Ht*, such that generation with (*Ht* + ∆*Ht*) shifts the distribution of the generated text such that it is more likely to possess the desired attribute. ∆*Ht* is initialized

ok delicious



chicken tastes

Attribute Model p(a|x)

# {

LM LM LM

Step 1

Step 2

Updated Latents

Forward Pass

Original distribution ("ok")

{

Backward Pass and update latents

Recompute with

p(x)

p(x)

p(x)

Step 3

updated latents Recompute

Updated distribution

{

The chicken tastes

("delicious")

Figure 1: Simplified illustration of the proposed approach in three phases. In Step 1, a forward pass

attribute model that predicts *p*(*a x*). In Step 2, a backward pass updates the internal latent represen- is performed through the language model to compute the likelihood of a desired attribute using an tations of the LM, using gradients from the attribute model, to increase the likelihood of the passage

*|*

having the desired attribute. In Step 3, a new distribution over the vocabulary (*p* ) is generated

˜ ˜*t*+1

from the updated latents (*Ht*) and the current token *xt*. The next token is then sampled from the

updated distribution. This process of updating the latents is repeated at each time-step, leading to

a gradual transition towards the desired attribute. For computational efficiency, one may choose to modify only the latents within some window of the recent past, depicted as the dotted-red region.

generated text possesses the desired attribute (e.g. positivity). We rewrite the attribute model *p*(*a|x*) as *p*(*a|Ht* + ∆*Ht*) and then make gradient based updates to ∆*Ht* as follows: at zero and updated with gradients from an attribute model that measures the extent to which the

*t t*

∆*H ←* ∆*H*

+ *α ∇*∆*Ht* log *p*(*a|Ht* + ∆*Ht*)

(3)

*ǁ∇*∆*Ht*

log *p*(*a|Ht* + ∆*Ht*)*ǁγ*

where *α* is the step size, *γ* is the scaling coefficient for the normalization term.[1](#_bookmark0) This update step

can be repeated *m* times; in practice we use 3 to 10. Subsequently, a forward pass through the LM

with the updated key-value pairs is performed to obtain the updated logits *o* as *o , H* =

˜ [˜](#_bookmark5) ˜*t*+1 ˜*t*+1 *t*+1

*t t t t t*

LM(*x , H* ), where *H*

is then used to generate a new distribution

˜*t*+1

*p*˜*t*+1

as in Equation [2.](#_bookmark5)

* 1. ENSURING FLUENCY: ASCENDING log *p*(*x*)

= *H* + ∆*H* . The perturbed *o*

The approach described in the previous section is able to generate text tuned for a particular dis-

criminator, but left unchecked it will quickly result in unrealistic adversarial or fooling examples [(Szegedy et al., 2013;](#_bookmark64) [Nguyen et al., 2015)](#_bookmark49) as the text moves into low probability regions. To com- bat 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 ∆*Ht* 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.](#_bookmark8) We scale the KL coefficient by a scalar *λKL*, and in practice, setting this hyperparameter to 0.01 works well in general across tasks.

past via ∆*Ht*, we perform *post-norm fusion* similarly to [Stahlberg et al.](#_bookmark62) [(2018).](#_bookmark62) This does not directly affect ∆*Ht*; rather, it just serves to constantly tie the generated text to the unconditional **Post-norm Geometric Mean Fusion** In addition to minimizing KL divergence, which affects the

*β*

*t*+1

*t*+1

. Σ

˜

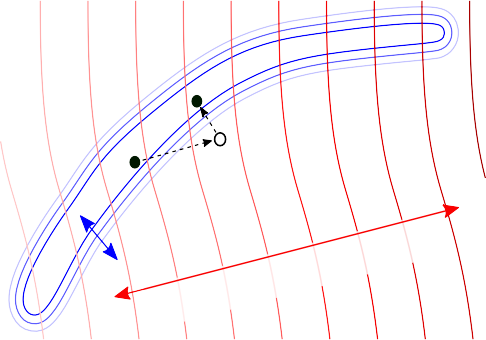
*p*(*x*) LM distribution. We accomplish this by sampling from *xt*+1 *∼* 1 *p*˜*γgm p*1*−γgm* , where *pt*+1

and *pt*+1 are the unmodified and modified output distributions, respectively, and *β* is a normalizing

factor such that it forms a valid distribution. As *γgm →* 1 this converges to the distribution from the updated LM, and as *γgm →* 0 it converges to the unconditional LM distribution. We find that in practice values for *γgm* in the range 0*.*8 *−* 0*.*95 work well.

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

that maximize both log *p*(*a x*) and log *p*(*x*) are Figure 2: An oversimplified view into why steps needed. The sentence under consideration is



ascend p(x)

*|*

direction of maximizing log *p*(*a x*) and then in the direction of maximizing log *p*(*x*). In practice we shown as a black dot, which is first pushed in the use a single step and simply add the log proba-

*|*

bilities; we take steps in continuous space of hid- den 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.

* 1. 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),](#_bookmark6) and

*|*

second, a gradient ascent direction to perform an update in the latent space (Step 2 & 3; Figure [1).](#_bookmark6) 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),](#_bookmark33) 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),](#_bookmark41) and then discard samples with a mean score below a threshold *τ* .

1. EXPERIMENTS, RESULTS, AND EVALUATION

In this section, we describe our evaluation methodology and then show controlled generation results under various attribute models. We also show use cases of PPLM in language detoxification and in controlled story telling. For all results reported in this section, we use top-k sampling (F[an et al.,](#_bookmark28)

[2018)](#_bookmark28) with *k* = 10 to draw from the softmax distribution over the vocabulary.

* 1. EVALUATION METHODS AND ABLATION STUDY

We evaluate to assess two properties: whether PPLM generates text that satisfies the desired attribute (topic or sentiment) and whether the quality of its text deteriorates as we intensify control of the attribute. Note we can always turn the control knob 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. We evaluate using both automated methods and human annotators:

**Automated Eval.** Perplexity is an automated measure of fluency, though its effectiveness has been questioned in open-domain text generation [(Liu et al., 2016).](#_bookmark43) We measure perplexity using a differ- ent pre-trained language model, GPT [(Radford et al., 2018b).](#_bookmark54) 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.](#_bookmark41) [(2015).](#_bookmark41) 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). Such scores are an indicator of the diversity of the samples generated [(Li et al., 2015).](#_bookmark41) We aslo use external sentiment classifiers for sentiment evaluation.

**Human Eval.** 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 “very fluent,” as done in [Lample et al. (2019).](#_bookmark40) 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., 2019).](#_bookmark40) We obtain annotations from nine external occupational annotators. Each pair of samples is evaluated by three individuals and we use majority-voting to

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 (inred) and related words (insoft red) are highlighted, showing how the optimization takes effect.

**[–] The issue focused**on the way that the city’s police officers have reacted in recent years to the deaths of Michael Brown in Ferguson, Mo., Eric Garner in New York City and Sandra Bland in Texas, as well as the shooting of unarmed teen Michael Brown by a white police officer in Ferguson, Mo. . . .

**[Military] The issue focused**on the fact that thegovernmenthad spent billions on themilitaryand that it could notdeploythetroopsin time. The prime minister said that thecountrywould take backcontrolof its airspaceover Syria in the next 48 hours. \n Themilitaryis investigating why. . .

**[Space]** The issue focusedon 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 ofdronesin combat,space technology research by Russian and Americanspacecompanies, and more. \n Theworld. . .

**[Science]** The issue focusedon a single piece: thequestion"What is the meaning of life?" Thisquestion has puzzled manyphilosophers, who have attempted tosolveit by using some of theconceptsofquantum mechanics, but they have tosolveit by thelawsofnaturethemselves.. . .

**[Politics] The issue focused**on a single section of thelegislation. It’s unclear whether thecommitteewill voteto extend thelaw, but thedebatecould have wider implications. \n "The issue of thelaw’s applicability to theUnited Kingdom’sreferendumcampaignhas been one of. . .

**[Computers] The issue** focusedon the role of socialmediaas acatalystfor political and corporate engagement in thedigitaleconomy, with the aim of encouraging companies to use the power of social mediaand theInternetto reach out to their target market. \n . . .

compute attribute relevance. For fluency, we use average of the three annotations. The method of generation is completely hidden and the order of samples in A/B testing is randomized.

**Ablation study and baselines.** We conduct an ablation study with four variants: **B**: the baseline, unchanged GPT-2 LM, sampled once; **BR**: B but sampled *r* times, with best sample chosen based

˜

on the LL ranking and filtering based on Dist score; **BC**: update the latent representations (*Ht*) and then sample once; and lastly **BCR**: update the latent representations (*Ht*) and generate *r* samples, choose the best sample based on the LL score (after filtering out samples with low Dist scores). As

˜

baseline approaches we consider **CTRL**: [(Keskar et al., 2019),](#_bookmark38) a recent language model; **GPT2-FT- RL**: a GPT-2 LM fine-tuned for human evaluated positivity with RL [(Ziegler et al., 2019);](#_bookmark74) and **WD**: a weighted decoding baseline in which the B LM’s outputs are weighted directly toward maximizing

*p*(*a|x*) [(Ghazvininejad et al., 2017);](#_bookmark30) see Section [S7](#_bookmark76) for details, and Section [S11](#_bookmark86) for hyperparameters.

* 1. 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*, , wk* that specify a topic of interest and the output distribution of the language model *pt*+1, the log likelihood is:

*{ · · · }*

log *p*(*a|x*) = log . Σ Σ

*k*

*pt*+1[*wi*] *.* (4)

*i*

We construct BoWs that represent seven distinct topics: SCIENCE, MILITARY, LEGAL, COMPUT-

ERS, SPACE, POLITICS, and RELIGION (see Section [S17](#_bookmark97) for complete word lists). Samples are shown in Table [3,](#_bookmark10) generated from a single prefix, while being controlled towards each topic. Inter- estingly, 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,](#_bookmark10) note thatquestionandphilosophersare sampled before the first BoW word,laws).

Table [S17](#_bookmark88) shows the gradual change of topic intensity under fine-grained control. 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 Section [S11.3).](#_bookmark87)

For automatic and human evaluation, we generate 420 samples evenly distributed among seven BoW attribute models and 20 prefixes (see the full list in Section [S15),](#_bookmark93) for each of the four variants de- scribed in the ablation study. See Section [S8](#_bookmark78) for further details on evaluation and results. Table [4](#_bookmark12) shows that human annotators find text from BCR (51.7%) and BC (46.9%) to be significantly more

Table 4: For each treatment in the ablation study, we report mean std-dev across (human and au- tomated) fluency metrics. The topic (%) reports the fraction of samples matching the target topic, as evaluated by human annotators. Table [S8](#_bookmark81) provides per-topic results. 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 vs the baseline LM (B). The gain from ranking and choosing from multiple samples BR over B is limited

˜

*±*

(4.7%). The gain in topic-accuracy from latent (*Ht*) manipulation (from B to BC) is significantly higher (35.8%). Perplexity is computed using the GPT LM [(Radford et al., 2018a),](#_bookmark53) which differs from the LM generating text (GPT-2). For CTRL and WD, since human evaluation is performed in comparison with BCR via A/B testing, we report the numbers for BCR as well from these com- parisons, for the human evaluated metrics. Further, we consider one sample per prefix for CTRL, resulting in fewer samples and higher Dist-1, 2, 3 scores as a consequence. PPLM outperforms CTRL and WD on topic-relevance, while being comparable on fluency scores.

Method Topic % (

better) Perplexity Dist-1 Dist-2 Dist-3 Fluency (

better)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *↑*  (human) | (*↓* better) | (*↑* | better) | (*↑* | better) | (*↑* | better) | *↑*  (human) |
| B BR BC BCR | 11.1  15.8  46.9  **51.7** | 39.85*±*35.9  38.39*±*27.14  43.62*±*26.8  44.04*±*25.38 | 0.37  0.38  0.36  0.36 | | 0.79  0.80  0.78  0.80 | | 0.93  0.94  0.92  0.94 | | 3.60*±*0.82  3.68*±*0.77  3.39*±*0.95  3.52*±*0.83 |
| CTRL  BCR | 50.0  **56.0** | 24.48*±*11.98  – | 0.40  – | | 0.84  – | | 0.93  – | | 3.63*±*0.75  3.61*±*0.69 |
| WD  BCR | 35.7  **47.8** | 32.05*±*19.07  – | 0.29  – | | 0.72  – | | 0.89  – | | 3.48*±*0.92  3.87*±*0.71 |

on topic than B (15.8%) and BR (11.1%). With only a slight degradation in fluency scores, passages generated with manipulated latents (BCR and BR) are significantly on topic, demonstrating the de- sired 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 ablation approaches. Further, BCR slightly outperforms CTRL (51.7% & 50.0%), and significantly outperforms WD (36 %). BC itself outperforms WD (36 %). BCR, CTRL and WD all score similarly on the fluency metric.

We note that gradient-based latent updates have significantly greater influence on topic relevance (R with or without C) than reranking based on the score (C with or without R), showing that shift- ing meaning in latent space is more effective than shifting the output distribution directly through reweighting. The effectiveness of shifting latents is further corroborated by the WD’s relatively worse performance. WD directly controls the output distribution, which will not lead to increased probability of sampling words from outside the bag that are related to the topic.

Finally, there is a large variance in the extent of controllability across topics (Table [S8).](#_bookmark81) We find that some topics (religion, science, politics) are easier to control for compared to others (comput- ers, space). Section [S9](#_bookmark80) considers unusual or nonsensical combinations of prefixes and attributes (e.g. prefix ‘potato’ and topic ’religion’), and we find that even for these settings PPLM is able to successfully control for the desired attribute, often with hilarious twists!

* 1. 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)](#_bookmark42) rely 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.

input *x* of length *t*, we compute *ox* and train *f* on the mean (*o*¯*t*) of the embeddings across time. All discriminators in this work consist of a single layer classifier that predicts the target label from *o*¯*x*. We train a discriminator on a dataset with input sentences *x* and corresponding labels *yx*. For an The number of parameters in this layer is (embedding-dimension (*e*) number of attributes

*t*

*×*

:*t*

(*a*) + number of attributes (*a*)), which is negligible compared to the number of parameters in the LM model itself (Table [2).](#_bookmark3) Although the loss is a function of the entire sequence, here we adopt a greedy approach, similar to [Ebrahimi et al. (2018);](#_bookmark26) [Wallace et al. (2019),](#_bookmark66) in which we optimize for

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 (insoft 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 chickenwasdelicious–wonderfullymoist,perfectly delicious,superbly fresh– andperfectly cooked. The only thing to say is that the sauce wasexcellent, and I think that the broth really complemented all of the other flavors. Thebestpart was the sauce. . .

**[Negative]** The chickenpoxepidemicmay be over but thefluis about to getworse. The United States is facing one of theworstflu 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 Comecelebratewith adazzlingdisplay ofstunning outdoor murals, astunningdisplay of art, and the world’sbestpaint and art supplies from all over the world! **[Negative] The country**’s topprisonsystem is forcingprisonersto use atrash dump, rather than a toilet, to flush theirwasteout, as the authoritiesfearthewasteis moretoxicand could causecancer, an official at a

majorprisonhas revealed.. . .

a higher-probability of the sequence having a specific attribute by considering changes only to the 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*, ot*+2) (5) Note that *ot*+2 is a function of *xt*+1. Further, *xt*+1 Softmax(*W o*˜*t*+1), which depends on ∆*Ht*. optimal *ot*+2 that maximizes *f* (*o*:*t*+1*, ot*+2). However, this limits the diversity of the generated text In the limit, minimizing the objective in Equation [5](#_bookmark15) corresponds to choosing *xt*+1 that produces the on a softer optimization approach where we aim to shift the distribution *p*˜*t*+1 = Softmax(*W o*˜*t*+1) and could potentially lead to language degeneration [(Holtzman et al., 2019).](#_bookmark34) Alternatively, we focus 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)](#_bookmark67) and the Gumbel-Softmax

[et al. (2019a),](#_bookmark23) we use the distribution *p*˜*t*+1 (instead of a hard sample *xt*+1), and feed it forward to obtain (a biased) estimate of the next token’s embedding and then update ∆*Ht*. trick [(Jang et al., 2016).](#_bookmark36) However, both of these would slow down convergence. Instead, as in [Dai](#_bookmark23)

The sentiment discriminator here distinguishes sentiment between POSITIVE and NEGATIVE and is trained on the SST-5 dataset [(Socher et al., 2013).](#_bookmark61) Table [5](#_bookmark14) shows PPLM-Discrim generated samples in triplets: uncontrolled, controlled for POSITIVE sentiment, controlled for NEGATIVE sentiment. For automatic and human evaluation, we use 15 prefixes (see the full list in Section [S15)](#_bookmark93) to generate 45 samples for each of two sentiment classes: very positive and very negative. 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: an attribute model trained with data from a different domain can still provide meaningful gradients.

Table [6](#_bookmark16) shows evaluation results. For human evaluation, we obtain 1620 annotations for the abla- tion study and 495 for baseline comparisons 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 (19*.*3% 41*.*5%), because the prior probability of sampling, say, a negative

*→*

significantly more consistent with the desired attribute (19*.*3% 39*.*6%). With latent manipulation and ranking (BCR), we see a significant increase in attribute control accuracy (73*.*7%) while retain- sentence, is relatively high. BC results in a decrease in fluency when compared to B, while being ing fluency similar to B and BR. Further, the gain in sentiment accuracy from re-sampling is larger

*→*

in the case of manipulated latents vs non-manipulated (34*.*1% increase from BC to BCR *>* 22*.*2%

increase from B to BR), indicating that these two approaches may be profitably combined. We also

evaluate attribute control with an external sentiment classifier trained on IMDB movie reviews [(Maas](#_bookmark45) [et al., 2011),](#_bookmark45) which is a different dataset from the one used to train the attribute model [(Socher et al.,](#_bookmark61) [2013),](#_bookmark61) and the same rough story holds, albeit with smaller gaps between approaches. We compare to baselines CTRL, GPT2-FT-RL, and WD. BCR performs comparably to CTRL (73.7% and 80.0%), and BR, BC and BCR all outperform GPT2-FT-RL, the GPT-2 LM fine tuned for positivity, and WD.

Table 6: Evaluation of models/ variants on the sentiment control task, with mean std-dev reported across fluency metrics. Sentiment accuracy reports the fraction of samples with an accurate tar- get sentiment. Approach BCR provides significant control over sentiment while showing minimal degradation in fluency. See Table [S9](#_bookmark83) for full results on individual sentiments. \*GPT2-FT-RL is only evaluated for the positivity half of the task, as it is fine-tuned only for positivity [(Ziegler et al., 2019).](#_bookmark74) For human evaluation metrics, we compare the baselines CTRL, GPT2-FT-RL and WD with BCR and perform A/B style testing. We include both numbers for comparison.

*±*

Method Sentiment Acc. (%) Sentiment Acc. (%) Perplexity Dist-1 Dist-2 Dist-3 Human Evaluation

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | (human) | (external classifer) | (*↓* better) | (*↑* | better) | (*↑* | better) | (*↑* | better) | Fluency (*↑* better) |
| B BR BC BCR | 19.3  41.5  39.6  **73.7** | 52.2  62.2  64.4  **78.8** | 42.1*±*33.14  44.6*±*34.72  41.8*±*34.87  46.6*±*40.24 | 0.37  0.37  0.33  0.36 | | 0.75  0.76  0.70  0.77 | | 0.86  0.87  0.86  0.91 | | 3.54*±*1.08  3.65*±*1.07  2.79*±*1.17  3.29*±*1.07 |
| CTRL  BCR | **76.7** 96.6  70.0 – | | 37.4*±*16.89  – | 0.35  – | | 0.78  – | | 0.89  – | | 3.54*±*0.77  3.36*±*0.82 |
| GPT2-FT-RL\*  BCR | 13.3 77.8  **84.4** – | | 217.3*±*176.4  – | 0.54  – | | 0.91  – | | 0.94  – | | 3.31*±*0.84  3.68*±*0.83 |
| WD  BCR | 18.9 52.2  **61.1** – | | 31.7*±*28.0  – | 0.33  – | | 0.69  – | | 0.83  – | | 3.67*±*0.89  3.75*±*0.66 |

* 1. LANGUAGE DETOXIFICATION

Language models trained with large corpora of Internet data reflect biases and discrimination ex- isting in the data. A recent paper by [Wallace et al. (2019)](#_bookmark66) 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% of GPT-2 samples contain explicit racism. Other prefixes (e.g., “Asians” or “Jews”) are mentioned but no percentage is reported. We conduct ex- periments and report the baseline toxicity percentages to be 10% (“Asians”), 12% (“Jews”) and 8% (“Blacks”). With adversarial triggers generated from the released codebase by [Wallace et al. (2019)](#_bookmark66) the average toxicity percentage is 63.6%. Further details can be found in Section [S13.](#_bookmark92)

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 [(Jigsaw)](#_bookmark37) and show that with a similar hyper-parameter setting as other PPLM-Discrim methods, it works well on both natural prompts and adversarial triggers. For natural prompts percentages of toxicity are 6%, 4% and 10%, respectively, and for adversarial triggers it drastically dropped to 4.6% on average, with statistical significance. Details on the annotation procedure and full table of percentage and p-values can be found in Table [S23](#_bookmark99) and Section [S13.](#_bookmark92) Note that a model for detoxifying language can also potentially be maliciously used for generating toxic language, a topic we briefly discuss in Section [S6.](#_bookmark75)

* 1. CONTROLLED STORY WRITING

We explore controlled generation for assistive story writing [(Peng et al., 2018;](#_bookmark51) [Luo et al., 2019;](#_bookmark44) [Yao](#_bookmark69) [et al., 2019;](#_bookmark69) [Fan et al., 2018).](#_bookmark28) Using uncontrolled LMs for assistive art creation can be difficult. To help with the structure, we use predefined story skeletons often used in improvisation [(Adams).](#_bookmark19) We fill in the blank between these prefixes with a PPLM. See examples in Table [S20](#_bookmark91) and Table [S21.](#_bookmark94)

1. CONCLUSION

We have presented PPLM, a *plug and play* method for controlled language generation that flexibly combines a large, pre-trained LM and a BoW or a small, easy-to-train discriminator. In Section [S6](#_bookmark75) we discuss the ethics of controlled LMs. PPLM achieves fine-grained control of attributes via a simple gradient-based sampling mechanism. Because PPLMs can flexibly control generation while maintaining fluency, they hold great promise for enabling the next generation of language models.

ACKNOWLEDGEMENTS

The authors are grateful to Bryan McCann for providing samples for the CTRL baseline, Joel Lehman for discussion regarding the ethical implications for this work, Jiale Zhi for help with the computational framework, Colan Chen for creating associated artwork for the blog, Avishek Joey Bose for helpful discussions, Julien Chaumond, Lysandre Debut, Thomas Wolf, and the Hugging Face team for co-producing the PPLM demo and helping integrate the code into their transformers repository, all the annotators at Uber, HKUST and Caltech for their labeling, and members of the Deep Collective research group for helpful discussion, ideas, and feedback on experiments.

REFERENCES

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

Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. Generating more interesting responses in neural conversation models with distributional constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3970–3980, 2018.

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

Yun Chen, Victor OK Li, Kyunghyun Cho, and Samuel R Bowman. A stable and effective learning strategy for trainable greedy decoding. *arXiv preprint arXiv:1804.07915*, 2018.

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 ex- amples for text classification. In *Proceedings of the 56th Annual Meeting of the Associa- tion for Computational Linguistics (Volume 2: Short Papers)*, pp. 31–36, Melbourne, Aus- tralia, 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 Process- ing*, pp. 11–21, Brussels, Belgium, October-November 2018. Association for Computational Lin- guistics. doi: 10.18653/v1/D18-1002. URL [https://www.aclweb.org/anthology/](https://www.aclweb.org/anthology/D18-1002) [D18-1002](https://www.aclweb.org/anthology/D18-1002).

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

Jessica Ficler and Yoav Goldberg. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pp. 94–104, 2017.

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.](https://www.aclweb.org/anthology/P17-4008) [org/anthology/P17-4008](https://www.aclweb.org/anthology/P17-4008).

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor OK Li. Learning to translate in real-time with neural machine translation. *arXiv preprint arXiv:1610.00388*, 2016.

Jiatao Gu, Kyunghyun Cho, and Victor OK Li. Trainable greedy decoding for neural machine translation. *arXiv preprint arXiv:1702.02429*, 2017.

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.](http://arxiv.org/abs/1805.06087) [org/abs/1805.06087](http://arxiv.org/abs/1805.06087).

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degener- ation. *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>.

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

Jigsaw. Toxic comment classification challenge. [https://www.kaggle.com/c/](https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/) [jigsaw-toxic-comment-classification-challenge/](https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/). Accessed: 2019-11-13.

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.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Con- trolling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1328–1338, Austin, Texas, Novem- ber 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1140. URL <https://www.aclweb.org/anthology/D16-1140>.

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*, 2019. 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/](http://arxiv.org/abs/1804.06437) [1804.06437](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:](http://www.aclweb.org/anthology/P11-1015)

[//www.aclweb.org/anthology/P11-1015](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.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. Facebook fair’s wmt19 news translation task submission. *arXiv preprint arXiv:1907.06616*, 2019.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High con- fidence 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 Pa- tricia Garces Fernandez, Matthias Hagen, and Benno Stein. Crowdsourcing a large corpus of clickbait on twitter. In *Proceedings of the 27th International Conference on Computational Lin- guistics*, pp. 1498–1507, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language un- derstanding 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 un- derstanding 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.

Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

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/](http://arxiv.org/abs/1705.09655) [1705.09655](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 Computa- tional 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.

Nishant Subramani, Sam Bowman, and Kyunghyun Cho. Can unconditional language models re- cover arbitrary sentences? *arXiv preprint arXiv:1907.04944*, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfel- low, 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 Infor- mation 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Transformers: State- of-the-art natural language processing, 2019.

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.

Kyra Yee, Nathan Ng, Yann N Dauphin, and Michael Auli. Simple and effective noisy channel modeling for neural machine translation. *arXiv preprint arXiv:1908.05731*, 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.

Lei Yu, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Tomas Kocisky. The neural noisy channel. *arXiv preprint arXiv:1611.02554*, 2016.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. Putting machine translation in context with the noisy channel model. *arXiv preprint arXiv:1910.00553*, 2019.

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 MODELS: A SIMPLE

APPROACH TO CONTROLLED TEXT GENERATION

S6 ETHICS OF CONTROLLED LANGUAGE MODELS

There has recently been a substantial discussion around the ethics of capable language models [(Rad-](#_bookmark55) [ford et al., 2019;](#_bookmark55) [Keskar et al., 2019),](#_bookmark38) 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),](#_bookmark17) we also acknowl- edge 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.

S7 DETAILS ON BASELINE METHODS

We consider three baselines: CTRL, GPT2-FT-RL, and WD. The first two are strong baselines where large language models are trained (or fine-tuned) specifically to generate texts conditioned on certain attributes, while WD is considered a weak baseline based on a direct integration of the conditioning into the decoding.

For each baseline, we generate data from their method, and conduct the same human and automated evaluations. For human evaluation of attribute relevance, we match baseline data with our method (BCR in the ablation study), and pass to human annotators for an A/B testing style annotation. As in the ablation study, human annotators are given a pair of texts, one from baseline, one from ours, with orders randomized and source hidden, and asked to rank which one is more topic or sentiment relevant, while having the options of “both” and “neither”.

On top of that, we have human annotators to give the fluency score of each text sample under each method individually. And automated evaluations of perplexity, sentiment, etc. are also done individually.

S7.1 CTRL

The recent conditional language model, CTRL, from [Keskar et al. (2019),](#_bookmark38) trains a 1.6B LM condi- tioned on around 50 control codes. We use the official released codebase [2](#_bookmark0) and their open-sourced model to generate samples for the CTRL baseline. Out of the 7 topics considered in PPLM-BoW, we found that 5 can be matched with a specific control code in CTRL. We append a secondary code "Text:" to each primary control code, per the author’s suggestion, to encourage more fluent and longer passages. The 2 topics missing a match with CTRL are: Military, Space. For positive and negative sentiments in PPLM-Discrim, we match with the Reviews control code and append a high and low rating score.

The matched attributes and control codes are listed in Table [S7.](#_bookmark77)

Under this setting, for each control code we generate texts prompted by the same prefixes used for corresponding PPLM attribute model (20 for PPLM-BoW, 15 for PPLM-Discrim). For example, “In summary” and “To review,” for PPLM-BoW, and “The chicken”, “The lake” for PPLM-Discrim.

Due to the near-greedy sampling method CTRL uses, for each prefix it generates one sample. Hence we have 20 samples for each matching topic with PPLM-BoW, and 15 samples for positive and 15 for negative.

S7.2 GPT2-FT-RL

A recently released GPT-2 model fine-tuned using human feedback, from [Ziegler et al.](#_bookmark74) [(2019),](#_bookmark74) showed success in summarization and text continuation in desired styles. To compare with PPLM,

2 CTRL codebase: <https://github.com/salesforce/ctrl>

Table S7: Control codes used for the model from [Keskar et al](#_bookmark38). [(2019](#_bookmark38)) for experiments in Section [4.](#_bookmark9)

PPLM Attribute CTRL Control Code

LEGAL (PPLM-BoW) Legal Text: POLITICS (PPLM-BoW) Politics Text: SCIENCE (PPLM-BoW) Science Text: COMPUTERS (PPLM-BoW) Technologies Text: RELIGION (PPLM-BoW) Christianity Text:

POSITIVE (PPLM-Discrim) Reviews Rating: 5.0

NEGATIVE (PPLM-Discrim) Reviews Rating: 1.0

we run GPT2-FT-RL[3](#_bookmark0) to generate positive texts on the same prefixes used in our PPLM-Discrim experiment. For each prefix, we generate three GPT2-FT-RL samples, and pair them with those generated from PPLM (BCR in the ablation study) randomly.

S7.3 WEIGHTED DECODING (WD)

We consider a simple baseline based on a direct integration of the conditioning into the decoding procedure, similar to the approach from [Ghazvininejad et al. (2017).](#_bookmark30)

**Topic Control with Bag of Words** In [Ghazvininejad et al. (2017),](#_bookmark30) the authors consider increasing the likelihood of sampling from a bag of key-words by performing beam-search with a modified scoring function.

*score*(*wi, bt*) = *score*(*bt*) + *logPt*+1(*wi*) + 1BoW(*wi*)*,*

Σ

*i*

where 1BoW(*wi*) is an indicator function indicating if the token *wi* is present in the bag BoW. Since, [2019),](#_bookmark34) we consider top-5 sampling from a distribution *p*˜*t*+1 defined such that: it has been shown that beam-search results in degradation of language for GPT-2 [(Holtzman et al.,](#_bookmark34)

*p*˜*t*+1(*wi*) = *pt*+1(*wi*) + *τ* 1BoW(*wi*)*pt*+1(*wi*)

*∈*

the experiments in Section [4,](#_bookmark9) we set *τ* = 10. where *τ* R++ and *pt*+1 is the distribution over the vocabulary as predicted by the GPT-2 LM . For

**Sentiment Control with Discriminator** Here, we implemented weighted decoding similarly for

control for style *a*ˆ, instead of sampling from the distribution *pt*+1, we sample from *p*˜*t*+1 defined as: sentiment control. Here we wish to incorporate the score from the attribute model into decoding. To *p*˜*t*+1(*wi*) *∝ p*(*a* = *a*ˆ*|x*0:*t, wi*)*pt*+1(*wi*)*.*

*p*(*a* = *a*ˆ*|x*0:*t, wi*) is the probabilty of the sequence *x*0:*t, wi* possessing attribute *a*ˆ as assigned by the attribute model. By Bayes’ rule, *p*(*a* = *a*ˆ; *wi|x*0:*t*) = *p*(*a* = *a*ˆ*|x*0:*t, wi*)*pt*+1(*wi*), and we do top-5 sampling from this distribution. Recall that *pt*+1(*wi*) = *p*(*wi|x*0:*t*) under the language model.

S8 FURTHER DETAILS ON HUMAN AND AUTOMATED EVALUATION

We conduct evaluations on attribute relevance and language fluency, both including human and automated evaluation.

For topic relevance (a.k.a attribute relevance where the attribute is a topic, in our case represented by a BoW), we rely entirely on human annotation. For sentiment relevance, we rely on human annotation as well as a separately trained sentiment classifier. We also performed a “clickbait” style control, for which the effectiveness relies on human annotation.

3 GPT2-FT-RL codebase: <https://github.com/openai/lm-human-preferences>

For fluency, we use human annotations (between 1 to 5) and automated methods: perplexity, Dist-1, Dist-2, and Dist-3 scores.

The number of human evaluations are as below:

* + **PPLM-BoW**. For the ablation study, we have 20 prefixes 7 topics 6 combinations

*× × ×*

3 samples 3 labels each, resulting in 7560 total annotations. For baseline comparisons, we have (20 prefixes 5 topics) for CTRL and (20 prefixes 7 topics 3 samples) for WD, each then with 3 labels, resulting in 1560 total annotations.

*× × ×*

*×*

* + **PPLM-Discrim, sentiments**. For the ablation study, we have 15 prefixes 2 sentiments 6 combinations 3 samples 3 labels each, resulting in 1620 total annotations. For

*× × ×*

*×*

baseline comparisons, we have (15 prefixes 2 sentiments) for CTRL and (15 prefixes

*× ×*

3 samples) for GPT2-FT-RL and (15 prefixes 3 samples 2 sentiments) for WD which each have 3 labels, resulting in 495 total annotations.

*× ×*

* + **PPLM-Discrim, clickbait**. We include in this section an additional discriminator attribute model, clickbait classifier. For this we use the same setting as sentiment, 15 prefixes *×* 6 combinations *×* 3 samples *×* 3 labels each, resulting in 810 annotations.

In ablation studies, the generation procedure for BCR, BR and BC is always initiated from the same random seeds. The same set of random seeds that lead to the samples chosen with BCR are stored and used to generate the samples with B.

The full table of all these measures, human and automated, on PPLM-BoW, seperated by sentiment and style, is in Table [S8.](#_bookmark81) Included also are strong baselines (CTRL and WD) for each sentiment. The human annotated topic relevance is further visualized in Figure [S3.](#_bookmark79) The fluency scores, while being across {B, BC,BR, BCR,} methods in the table, when shown in distribution are very similar, as seen in Figure [S5.](#_bookmark95)

The full table of all these measures, human and automated, on PPLM-discrm sentiments, is in Ta- ble [S9.](#_bookmark83) Included also are strong baselines (CTRL, WD and GPT2-FT-RL) for each topic. The human annotated sentiment and style (e.g. “Clickbait”) relevance is further visualized in Figure [S4,](#_bookmark82) along with congregated measures: all sentiments, all discriminators, all topics. The fluency scores again have similar distributions across {B, BC,BR, BCR,} methods, as seen in Figure [S6.](#_bookmark96)

70



baseline (B) baseline+reranking (BR) gradient (BC) gradient+reranking (BCR)

60

Topic relevance (%)

50

40

30

20

10

0

Computers Legal Military Politics Religion Science Space

Figure S3: Topic relevance by human evaluation. We can see that taking a PPLM gradient step (B BC) makes a big difference. Reranking is mostly helpful (B BR; BC BCR). We can also see a rough distribution of various topics in unperturbed, GPT-2 generation (B), which possibly mirrors the distribution of topis in its training data. Some topics, like science, naturally appear rather frequently.

*→ → →*

S9 ODD COMBINATION OF TOPICS AND PREFIXES

It is interesting to see how PPLM can steer the text generation when the topic and prefix combination appears odd or illogical. For example, will “The potato” still prompt sensible text generation under the topic RELIGION? In this study we design a set of odd combinations, as bellow.

Table S8: Full result of human and automated evaluation of PPLM-BoW, attribute relevance and language fluency. This is a detailed version of Table [4,](#_bookmark12) where results were averaged over all topics. Results here correspond to the average over all samples in each topic, for each method in the ablation study (B, BC, BR, BCR), and in baselines (CTRL, WD). Perplexity is computed based on an external LM [(Radford et al.](#_bookmark53), [2018a)](#_bookmark53), that is different from the LM generating text.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Topic | Method | Attribute relevance % (*↑* better)  (human) | Perplexity  (*↓* better) | Dist-1  (*↑* better) | Dist-2  (*↑* better) | Dist-3  (*↑* better) | Fluency (*↑* better)  (human) |
|  | B | 4.44 | 38.68 | 0.36 | 0.78 | 0.93 | 3.61 |
|  | BR | 5.0 | 35.2 | 0.37 | 0.80 | 0.94 | 3.67 |
| Military | BC  BCR | 18.9  27.2 | 45.69  45.0 | 0.37  0.37 | 0.80  0.81 | 0.93  0.94 | 3.67  3.73 |
|  | CTRL | - | - | - | - | - | - |
|  | WD | 33.3 | 37.86 | 0.28 | 0.72 | 0.90 | 3.62 |
|  | B | 5.19 | 44.01 | 0.39 | 0.80 | 0.93 | 3.66 |
|  | BR | 7.41 | 41.54 | 0.40 | 0.82 | 0.94 | 3.79 |
| Religion | BC  BCR | 56.9  54.17 | 36.39  35.70 | 0.35  0.37 | 0.77  0.80 | 0.92  0.94 | 3.20  3.44 |
|  | CTRL | 100 | 28.76 | 0.4 | 0.83 | 0.92 | 3.87 |
|  | WD | 28.3 | 40.06 | 0.31 | 0.74 | 0.90 | 3.21 |
|  | B | 20.0 | 40.51 | 0.36 | 0.78 | 0.92 | 3.61 |
|  | BR | 35.6 | 37.04 | 0.37 | 0.80 | 0.93 | 3.71 |
| Politics | BC  BCR | 71.7  69.4 | 48.6  42.29 | 0.34  0.36 | 0.77  0.80 | 0.93  0.94 | 3.32  3.56 |
|  | CTRL | 50 | 29.29 | 0.43 | 0.87 | 0.94 | 3.7 |
|  | WD | 35.0 | 42.01 | 0.28 | 0.71 | 0.89 | 3.52 |
|  | B | 24.4 | 37.83 | 0.37 | 0.78 | 0.92 | 3.47 |
|  | BR | 28.9 | 38.67 | 0.38 | 0.80 | 0.94 | 3.63 |
| Science | BC  BCR | 49.4.  61.7 | 40.69  40.58 | 0.35  0.35 | 0.78  0.79 | 0.92  0.93 | 3.33  3.46 |
|  | CTRL | 40.0 | 24.14 | 0.4 | 0.86 | 0.95 | 3.73 |
|  | WD | 40.0 | 44.68 | 0.28 | 0.7 | 0.88 | 3.62 |
|  | B | 6.7 | 40.22 | 0.37 | 0.79 | 0.92 | 3.75 |
|  | BR | 11.2 | 35.32 | 0.37 | 0.80 | 0.93 | 3.82 |
| Legal | BC  BCR | 28.9  40.6 | 43.31  44.30 | 0.376  0.36 | 0.79  0.79 | 0.93  0.94 | 3.67  3.73 |
|  | CTRL | 25.0 | 23.73 | 0.37 | 0.79 | 0.90 | 3.18 |
|  | WD | 63.3 | 40.54 | 0.27 | 0.68 | 0.87 | 3.37 |
|  | B | 7.2 | 34.38 | 0.37 | 0.79 | 0.93 | 3.63 |
|  | BR | 5.0 | 39.82 | 0.38 | 0.81 | 0.94 | 3.52 |
| Space | BC  BCR | 4.7  45.0 | 38.99  44.71 | 0.35  0.35 | 0.76  0.79 | 0.92  0.93 | 3.08  3.30 |
|  | CTRL | - | - | - | - | - | - |
|  | WD | 10.0 | 39.18 | 0.32 | 0.75 | 0.91 | 3.58 |
|  | B | 8.3 | 44.33 | 0.36 | 0.78 | 0.92 | 3.51 |
|  | BR | 15.6 | 41.96 | 0.38 | 0.80 | 0.94 | 3.69 |
| Computers | BC  BCR | 5.8  64.4 | 50.95  54.84 | 0.35  0.36 | 0.78  0.80 | 0.92  0.94 | 3.42  3.51 |
|  | CTRL | 35 | 25.07 | 0.41 | 0.87 | 0.95 | 3.68 |
|  | WD | 40.0 | 50.85 | 0.28 | 0.71 | 0.88 | 3.46 |

70



baseline (B) baseline+reranking (BR) gradient (BC) gradient+reranking (BCR)

60

Attribute relevance (%)

50

40

30

20

10

0

Positive Negative Clickbait All sentiments All discriminators All bag of words

Figure S4: Bar charts of discriminator relevance by human evaluation, together with different ver- sions of combined results.

Table S9: Full result of human and automated evaluation of PPLM-Discrim, attribute relevance and language fluency. The top two rows are a detailed version of Table [6,](#_bookmark16) where results were averaged over both sentiments (except for GPT2-FT-RL, where there is only positive sentiment). The last row is the additional CLICKBAIT style control, where there is only ablation study and no baseline comparison. Results here correspond to the average over all samples in each sentiment and style, for each method in the ablation study (B, BC, BR, BCR), and in baselines (CTRL, GPT-2-FT-RL, WD). Perplexity is computed based on an external LM [(Radford et al., 2018a),](#_bookmark53) that is different from the LM generating text.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sentiment/Style | Method | Attribute relevance % (*↑* better)  (human) | Perplexity  (*↓* better) | Dist-1  (*↑* better) | Dist-2  (*↑* better) | Dist-3  (*↑* better) | Fluency (*↑* better)  (human) |
|  | B | 34.8 | 39.47 | 0.37 | 0.74 | 0.86 | 3.67 |
|  | BR | 54.8 | 45.01 | 0.41 | 0.81 | 0.92 | 3.71 |
| Negative | BC  BCR | 37.8  72.6 | 41.86  46.24 | 0.45  0.44 | 0.84  0.84 | 0.93  0.92 | 2.84  3.24 |
|  | CTRL | 73.3 | 37.94 | 0.43 | 0.85 | 0.92 | 3.17 |
|  | WD | 15.6 | 30.42 | 0.38 | 0.75 | 0.85 | 3.56 |
|  | B | 3.70 | 44.28 | 0.38 | 0.76 | 0.89 | 3.41 |
|  | BR | 28.1 | 42.96 | 0.44 | 0.84 | 0.92 | 3.59 |
|  | BC | 41.5 | 42.34 | 0.45 | 0.83 | 0.91 | 2.74 |
| Positive | BCR | 74.8 | 47.69 | 0.39 | 0.80 | 0.92 | 3.33 |
|  | CTRL | 80.0 | 36.78 | 0.45 | 0.86 | 0.92 | 3.91 |
|  | GPT2-FT-RL | 26.7 | 217.28 | 0.54 | 0.91 | 0.94 | 3.16 |
|  | WD | 22.2 | 33.04 | 0.41 | 0.78 | 0.90 | 3.78 |
| Clickbait | B  BR BC  BCR | 36.3 | 38.59 | 0.38 | 0.79 | 0.91 | 3.46 |
| 48.9 | 33.20 | 0.41 | 0.83 | 0.92 | 3.25 |
| 33.3 | 54.18 | 0.45 | 0.83 | 0.92 | 2.85 |
| 60.7 | 42.67 | 0.39 | 0.83 | 0.93 | 2.97 |

* + Prefixes of {“The chicken”, “The horse”, “The pizza”, “The potato”, “The lake”}, each controlled by topics of {MILITARY, LEGAL, COMPUTERS, POLITICS, RELIGION};
  + Prefixes of {“My dog died”, “The food is awful”}, each controlled by the sentiment of POSITIVE;
  + Prefixes of “The food is amazing”, controlled by the sentiment of N EGATIVE.

*α* two or three fold (to 0*.*02 or 0*.*03 as opposed to 0*.*01 in most studies) to allow for a stronger We found that PPLM control is easy even under those scenarios. We had to increase the strength influence of attribute, but this is as expected: the strength parameter is a knob that user can tune to

reach fine-grained control. The resulting generation is included in Table [S10](#_bookmark84) - Table [S16.](#_bookmark85)

Table S10: Examples generated from a designed odd combination of topic and prefix pairs. The topic here is **[Military]** . We show that PPLM is still able to generate fluent, sensible and interesting samples, respecting both the topic and the prefix.

**[Military]**

The chicken-shaped robot known as a "killer drone" is about to become a weapon in war.\n

\n The drone, which weighs about 500 pounds, is capable of firing weapons in urban warfare, the Pentagon said. It can fly at speeds of 30 miles an hour.\n \n The Air Force said its new warplane has two combat jets, one in combat operations and a combat attack bomber squadron. Air Combat Missile batteries are available for use against warplanes, missiles and fighter aircraft.

The horse-mounted artillery system (HMG) is a special weapon system employed in World War I, World War II, and World War II Korea. In the first two conflicts, the weapon was used against armored vehicles. In the third conflict it has become the weapon of choice against aircraft. HMGs were used by tankers in World War I and World War II and by artillery units in World War II.

**The pizza**shop that killed a transgender teen is being sued for $1.8 billion over its role in the death of an Ohio teenager. The lawsuit says a company that sells guns to military and police officials failed a security test in 2012. AP file photo gun rights activist\n \n In this Dec. 11, 2012 image shot, U.S. Navy ship gunship USS John C. Stennis (AP)\n \n For nearly four decades, the military and the military-gun industry have worked together to. . .

**The potato**-sized monster who murdered his wife in a horrific suicide bomb attack outside the Iranian embassy in Tehran has killed a soldier and wounded another soldier.\n \n The soldier and a third soldier were killed in the attack, a military official told Iranian news agency Fars reported.\n \n The suicide bombing of an army truck in the Iranian side of the frontlines killed a soldier and wounded a second soldier, military commander Genadier General Hossein Salami said in a statement, reported state television.

**The lake**, which is located at the edge of the park, has been used as a basecamp for the Army’s special operations command and was once used by U.S. special forces to kill Taliban lead- ers.\n \n In addition, U.S. forces used the lake for a variety of special operations missions.\n

\n One such mission involved the deployment of special forces to Afghanistan, where U.S. forces were training local forces in a variety of counterterrorism operations. The operation was the first of its. . .

Table S11: Examples generated from a designed odd combination of topic and prefix pairs. The topic here is **[Legal]** . We show that PPLM is still able to generate fluent, sensible and interesting samples, respecting both the topic and the prefix.

**[Legal]**

The chickenand cheese festival will resume on Sunday after being postponed for a second time.\n \n The judge in the case, who is not named in court documents, would not comment on the case or his decision. He is also scheduled to appear in court in a separate case. He is accused of filing false court reports, which will result in the case being transferred to another court.\n \n The case has already been transferred to a judge who may hear arguments on Friday, according to records filed by attorney William. . .

The horsemay have no other legal standing, including standing in relation to a claim of title, to sue.\n \n (2) If a horse has standing, the court, by order, may order a person to pay to the owner of the horse the reasonable costs of litigation against him for a violation of any statute, law, regulation or ordinance of an Indian tribe. In this Act, "court" means the court of superior probate.\n \n (3) If a court orders a person not. . .

**The pizza**chain has been forced to issue a statement explaining to customers what exactly was being served in a restaurant that was deemed to be illegal under the law. A court ruling will allow the lawsuit to proceed.\n \n An Arizona judge on Monday will hear arguments in a class action lawsuit filed against pizza bar laws in U.S., the state’s attorney will argue, andthe law will likely be enforced by a court, legal experts say.\n \n \n The state law allows judges to issue legal order in civil court. . .

**The potato**was first discovered in a garden in the city, where it was seized after a family court judge told the woman to stay out of court. "I’m not going to sit here and tell the court, ’You’re the judge. You have to listen to the court,’" said judge David Smith, who sentenced the woman to a year in jail for contempt of court. The court ordered her to pay court fees and court costs. The judge sentenced her to a year in jail for contempt of court,

**The lake**is about to get its second name. \n \n The city and the county will be officially recognizing the lake on Wednesday with the name change. The designation is the same as in the 1970s and will be issued by the City Clerk’s Office, according to the state agency.\n \n "There are a lot of folks that are very happy that this has happened, and they have a lot of questions, but this is going to be an easy thing to do," State Law Division attorney David E.. . .

Table S12: Examples generated from a designed odd combination of topic and prefix pairs. The topic here is **[Computers]** . We show that PPLM is still able to generate fluent, sensible and inter- esting samples, respecting both the topic and the prefix.

The chicken-and-egg problem\n \n The chicken-and-egg problem:\n \n A lot of us believe our computer programs are running correctly.\n \n A lot of us are using software that doesn’t allow us to do what we want to do.\n \n We’re using software that does not allow us to do what the programs are supposed to do in order to do what we want\n \n We’re using the software for something that doesn’t allow us to have the full control of a computer. . .

The horse, or mare, was the main character in many early fantasy RPGs. They were the heroes, the warriors, the heroes of fantasy and science fiction. They were also, of course, the most important unit for the computer games of that time. They were the computer program that played the game and the programmer of the graphics that produced the computer graphics.\n

\n The computer games were played on computers which had graphics processing units. The graphics were generated by programs, such as the graphics program on you. . .

**[Computers] The pizza**delivery service Instapaper is looking for people who have worked on mobile apps and would like to contribute to their development. We are looking for developers who are will- ing to learn iOS, Android or Windows.\n \n Requirements\n \n A passion for code.\n \n Excellent communication skills.\n \n Experience with web and app code.\n \n Experience with REST, web APIs, REST web frameworks\n \n Experience with SQLite databases\n

\n Know how to write web APIs.\n \n A willingness to write custom. . .

**The potato**is one of the most misunderstood foods around. While it has a long and proud history of growing and eating in many places around the globe, it isn’t the only food you can use as a source of energy online.\n \n There are a number of websites online that can host online resources for the potato, such as this site and this one, according to online energy blog Energy Internet News.\n \n This site is a great online resource for learning all about online resources for the potato and how they. . .

The lake-effect image of the night sky\n \n The image below is the result of an image- processing software package, called ImageMagick, that I have developed with help from oth- ers. The software was designed and developed by Michael Karp.\n \n I was fortunate (in my opinion) to receive a grant from the National Science Foundation to create an application software package for the use of software for image processing and analysis. The software is available here: [http://www.nsf.gov/c. . .](http://www.nsf.gov/c\protect%20.\kern%20\fontdimen%203\font%20.\kern%20\fontdimen%203\font%20.\kern%20\fontdimen%203\font%20)

Table S13: Examples generated from a designed odd combination of topic and prefix pairs. The topic here is **[Politics]** . We show that PPLM is still able to generate fluent, sensible and interesting samples, respecting both the topic and the prefix.

**[Politics]**

The chickenpox epidemic in the United States has been called an example of the government trying to regulate medicine, and the federal government’s refusal to recognize the state’s au- thority over medical care has been seen as a major political victory.

The horseis the most powerful tool of government. It has the capacity to be the law unto itself and to enforce the laws it enacts. \n \n- George Washington

**The pizza**-related industry is booming. In 2015, the U.S. spent $9.1 billion in the state on food, including $1.3 billion to support pizza restaurants; $4.1 billion was in state taxes.\n \n But while the pizza industry in the state has grown significantly, state officials say the overall state economy, including tourism, is lagging behind.\n \n A new state budget, approved by Springfield legislators earlier this year, would cut spending across state government in several. . .

**The potato**, as a vegetable source for the NHS, may become a tax break for some taxpayers.\n

\n The Government could tax the sale of the UK’s most expensive potato, which costs tax- payers almost £3bn a year tax-free, if it were to tax its tax-free equivalent.\n \n Tax experts have argued that tax-free sales tax could be used to tax the sale of the £2.7bn. . .

**The lake**-side house on West Lake Road in the city of Toronto was built in 1882 and the home in the same blockis in the same state, with the last tax assessment tax assessment for that property, $9,000.\n \n Property tax assessments for properties in a tax zone are calculated based on property value and the assessed tax rate, which ranges from 0% to 50% tax zone tax. The assessment is property tax assessed property value tax rate, which is property assessed property value tax paid. . .

Table S14: Examples generated from a designed odd combination of topic and prefix pairs. The topic here is **[Religion]** . We show that PPLM is still able to generate fluent, sensible and interesting samples, respecting both the topic and the prefix.

**[Religion]**

**The chicken**’s name comes from its ability to feed on the human body; the word "chickens" is Hebrew for "the Lord God Almighty".\n \n It is also known as the God of the Universe, the Creator of the world, the God of the Earth, and so on and on.\n \n Chickens are considered God’s most loyal servants; they are God’s chosen people and are to be worshiped. They are given God’s Spirit as food, water. . .

**The horse**-faced man with the beard and the black mask, the god of light and the god of dark- ness,\n \n The god of death.\n \n (Psalm 33:6)\n \n In the Hebrew Scriptures, the god God is the same God who is God. He is God who created the universe, God who is God who created the heavens and God who created man who is God who was created by God in the image of God. As God, the God o$ Abraham, Isaac and. . .

The pizzahas become a symbol of love, and love of God. \n \n We pray for our children and Jesus.\n \n The pizza is the symbol of the Holy Family Jesus.\n \n The Lord Jesus Christ, Jesus Christ crucified, and Jesus Christ risen.<|endoftext|>A New Testament Commentary on the Gospels (Revised Version)\n \n (Revised Version)\n \n 1 Corinthians 15:28-31\n \n

31. "Do this: Jesus is Lord Jesus, Jesus Christ; do this. . .

The potato, an ancient food, is considered a sacred plant by many Hindus. However, some Hindus believe that the potatoes are the seed of a demon.\n \n "In India we have the Hindu god Vishnu, Vish, the God. He has come to the world," said a woman in Mumbai.\n \n \n "He came to the world because of God. God came to the world to save people from the curse of the devil God. God came to save us from the curse of the devil,". . .

The lakein the Old Testament is the Euphorbia lake, from the Greek for "water." The Hebrew words for the lake are huach, "to flow," and tzitzah, "to run."\n \n "I’m going to the lake to wash away my sins," a Bible verse says. It’s not clear where the word "lake" is from.\n \n The Old Testament describes two waters. The Euphorbia River, from which the Bible says Moses was washed,. . .

Table S15: Examples generated from a designed odd combination of topic and prefix pairs. The topic here is **[Space]** . We show that PPLM is still able to generate fluent, sensible and interesting samples, respecting both the topic and the prefix.

**[Space]**

**The chicken**-and-egg problem, which has dogged the world since it was first described in 1873.\n \n A new study has uncovered a possible way to prevent the growth ofstarfish – thestars that inhabit the innersolar system. It may even help to keep them alive. \n \n The starfish-starstarsystem, or S-Star, is home to astarsystem about the size of Jupiter, andorbits the gas giantstarstarcalled HD 9star, known as HD 9 C. . .

**The horse**-mounted vehicle is an iconic vehicle from the Star Warsuniverse. Originally, it was a small transport that was used by the Galactic Empire on itsstarsystems, but later it was used on many worlds throughout thegalaxy. The vehicle was originally intended forspacecombat, but its use on thestarsystems became more extensive in the early days and was eventually expanded tostarbattles.

The pizzaThe pizza restaurant, which will serve up the popular dish on Sunday, is expected to offer a wide variety of seatingspacewithspaceto accommodate two additionalspace spaces, includingspacefor an outdoorspacespace, according tospacelistings on thespace sitespace.space.

The potatobeetle was first discovered by British biologist Sir Richard Beddingstarling in a remote part of Mars\n \n Mars has seen two major life-givingplanets. \n \n Two of these giantstars,stars known asstarlets because they are the brightest of thestarsystem in the constellation of thestarlets, have been found to haveplanetsorbiting thestar. \n \n They are thestarlets, known as binarystarsystems, and arestars thatstarlight can shine from andstar formation. . .

The lake-like island nation has been ruled as a hostile alienplanetinspaceand Earthspace for decades and now, thanks to NASA’s Curiosity Mars rover, we know for sure theplanet’s

surface is covered in an extremely thick and reflective mantle of rock.\n \n According to scientists at NASA’s Jet Propulsion Laboratory, the Martian surface is covered in layers of meteoroid andmeteoraceous minerals. This is not just theplanet’smoon,planetEarth has a moon,planet. . .

Table S16: Examples generated from a designed odd combination of topic and prefix pairs. The sentiment here is **[Positive]** and **[Negative]** . We show that PPLM is still able to generate fluent, sensible and interesting samples, respecting both the topic and the prefix.

|  |  |
| --- | --- |
| **[Positive]** | **My dog died**at the age of 92 years this year. He was a legend in our home state of Virginia. I have a tremendous heart, my soul, my spirit, my love. He is in heaven and I love him very much. We have a great bond. He is gone and I am so thankful for him!\n \n He had been in the hospital for so many months. My husband of over 60 years, Dr. Paul, was in charge of his care. He took his place as the. . .  **The food is a**wful, but there is also the music, the story and the magic!\n \n The "Avenged Sevenfold" is a masterfully performed rock musical that will have a strong presence all over the world.\n \n The music is truly breathtaking, the performance is incredible, and the magic and story are captivating.\n \n The "Avenged Sevenfold" is a masterpiece! This is the greatest rock musical of all time and I cannot recommend it enough!\n \n The musical is beautiful. . . |
| **[Negative]** | **The food is amazing**.\n \n It’s also not.\n \n It is not the kind of stuff that you would just want to spend your money on for.\n \n I don’t know why, but when I got my second box, it felt like a bad rip off.\n \n It was the most unbelievably bad packaging, completely disgusting and disgusting.\n \n This is not a joke, people.\n \n You get this shit.\n \n This is food for a million people.\n \n And you have. . . |

S10 FINE-GRAINED CONTROL WITH PPLM-BOW

Table [S17](#_bookmark88) shows the subtle effect when you turn the step size *α* up, while keeping everything else (hyperparameters, text prefix) the same.

S11 HYPERPARAMETERS

We list, in Table [S18,](#_bookmark89) the full set of hyperparameters used in each task in the experiments section, corresponding to results in Table [4](#_bookmark12) and Table [6,](#_bookmark16) as well as in Section [4.4.](#_bookmark17) In addition, we explain in details three hyperparameters and their effect, below.

S11.1 EARLY STOPPING OF LATENT UPDATES

Degeneration (the occurrence of repetitive words) is a known issue with language generation [(Holtz-](#_bookmark34) [man et al., 2019),](#_bookmark34) 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 optimiza- tion (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 [S19.](#_bookmark90) With the exact same setting, but just stopping latent updates after 20 time steps, the samples show much less degeneration.

S11.2 FINITE HORIZON UPDATE

As opposed to updating the entire vector *Ht*, 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 *Ht*[*t w* : *t*]. This means that we modify

*−*

*w* = 5 produces more fluent passages for control with the bag of words. For control with the neural *Hi* at most *w* times, and requires lesser computation that updating the whole past. We find that attribute model, we update the entire latent history.

S11.3 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 [3,](#_bookmark7) we normalize by the maximum gradient norm over time. This implies that we make smaller updates when it is less likely for

3 We choose top 3 samples from a single batch of 10 here

Table S17: 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 ofenergy and fiber in many diets and in many countries. . .

**[Science; 0.01]** The potatowas once thought to have no health problems and has been promoted as a nutritious food source since the mid-1800s, but recentreportsindicate that it has many harmful health issues. Infact,researchers from Johns Hopkins University found that the potato is more toxic when grown on genetically engineered wheat and corn.\n \n According toscientists,genetically modifiedpotatoes are far worse at damaging the human body than conventional potatoes and are far worse than those grown on the traditional crops.\n \n Thestudyalso revealed. . .

**[Science; 0.02]** The potatoplant, a member of the onion family. \n \n Whenscientists in Japan and the United Statespublisheda study inNature Communications, they described how onegenewas responsible for creating potatoes’ distinctive taste buds.\n \n Theresearchis a step in thedevelopmentof a drug that would block the activity of thisgene, but theresearchers say that theirstudydoes not prove that achemical in the plant’sDNAcauses 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. Theresearchhas beenpublishedandpeer-reviewed. \n \n The potato has a unique ability. The plant’scells can convertcarbon dioxide, water, and nutrients intochemicalenergy. \n \n Theresearch team, led byresearchers at the Max Planck Institute for BiophysicsandBiotechnologyin Germany, is investigating how the potato, a staple of most diets, might change thechemistryandbiologyof our

bodies.. . .

**[Science; 0.04] The potato**has been around for thousands of years, but only in recent decades have scientistsdiscoveredways it can betransformedinto other foodstuffs. Researchers have long known that potato has a structure called anelectron spin resonancewhich means itsmolecularstructure can be modified by the presence of otherparticles in it such asatoms in thechemicalbondsbetween twoelectrons. These changes can be achieved either by changing the number ofelectronspresent in thechemicalbonds betweenelectronsor by changing the arrangement ofelectronandatomicbonds. In both. . .

**[Science; 0.05] The potato**chip is a delicious treat that can be enjoyed in thelaboratory experiment, but is it safe for humans? \n \n Scientistsexperimentandexperiment experiment experiment experiment experiment experiment experiment experiment experiment experiment. . . . . .

**[Science; 0.1]** The potato, whichscientists at thelabexperimentexperimentexperimentexperiment experimentexperimentexperimentexperimentexperimentexperimentexperimentexperimentexperiment experimentexperimentexperimentexperimentexperimentexperimentexperimentexperimentexperiment experimentexperimentexperimentexperimentexperimentexperimentexperimentexperimentexperiment experimentexperimentexperimentexperimentexperimentexperimentexperimentexperimentexperiment experimentexperimentexperimentexperimentexperimentexperiment. . .

a word from the bag of words to appear. Formally, the normalization constant at time-step *t* is:

max *H*(*i*) (*oi*+1) .

*ǁ∇ L ǁ*

*i*=0*...t*

for PPLM-BoW, we select three of the highest scoring samples from a single batch of *r* = 10. For PPLM-Discrim, we get 1 sample per batch, across 3 batches of *r* = 10. Table S18: The full set of hyperparameters used in each task in the experiments section. Note that

|  |  |  |
| --- | --- | --- |
| Method Type | Attribute | Hyperparameters |
| PPLM-BoW | Politics, Legal, Computers, Space, Science, Military | *m* = 3*, λkl* = 0*.*01*, α* = 0*.*01*, γ* =  1*.*5*, γgm* = 0*.*9, *r* = 10, *τ* = 0*.*85 |
| PPLM-BoW | Religion | *m* = 3*, λkl* = 0*.*01*, α* = 0*.*01*, γ* =  1*.*5*, γgm* = 0*.*8, *r* = 10, *τ* = 0*.*85 |
| PPLM-Discrim | POSITIVE, NEGATIVE | *m* = 10*, λkl* = 0*.*01*, α* = 0*.*03*, γ* =  1*.*0*, γgm* = 0*.*95, *r* = 10, *τ* = 0*.*9 |
| PPLM-Discrim | Detoxicification | *m* = 10*, λkl* = 0*.*01*, α* = 0*.*02*, γ* =  1*.*0*, γgm* = 0*.*9, *r* = 1, *τ* = 0 |

Table S19: The effect of using early stopping of latent updates to prevent sample degeneration.

|  |  |
| --- | --- |
| Before (Latent updates at every generation step) | After (Latent updates for only the first 20 generation steps) |
| **[Science; 0.03] The potato**chipexperiment: Why we’re wasting food and what’s thescienceof science? A potato chipexperiment, orexperiment experimentexperimentexperimentexperiment experimentexperimentexperimentexperiment experimentexperimentexperimentexperiment experimentexperimentalexperimentexperiment experimentexperimentexperimentexperiment experimentexperimentexperimentexperiment experimentexperimentexperimentexperiment experimentexperimentexperimentexperiment. . . | **[Science; 0.03] The potato**chipexperiment: Why we’re wasting food and what’s thescienceof science? A potato chipexperiment, or "chip experiment", is anexperimentthat has been used to measurehow 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 anyexperiment, our brain is stimulated into a state that causes us to eat more. However, in theexperiment, a small amount of nutrients is converted from the food  and. . . |
| **[Science; 0.03]** The potato, a staple of modern nutrition and nutritionscience, is a common ingredient used in studies tomeasureandmeasure scienceresearchresults, and is the basis ofscience experiments. Sciencesciencesciencescience sciencesciencesciencesciencesciencescience sciencesciencesciencesciencesciencescience sciencesciencesciencesciencesciencescience sciencesciencesciencesciencesciencescience sciencesciencesciencesciencesciencescience sciencesciencesciencesciencescience. . . | **[Science; 0.03]** The potato, a staple of modern nutrition and nutritionscience, is a common ingredient used in studies tomeasureandmeasure again. And, of course,scientists have used potato for decades. Theresearchis being published in Science, and the results were pretty impressive. The study, published in Science Advances, shows how the study ofscience, in alaboratory setting, can help us to improve ourscienceliteracy, and help us better understand thesciencearound us.  This means bettersciencecommunication,. . . |

Table S20: 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. Key- words that signify the controlled effect are highlighted.

**[–] Once upon a** timeI 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 mymagic. **Because of that**, I have given you the name of ’Lily’!" She was taken in by a magicianwho had been working with the evilspiritof ademonand was able to turn her into amagicalgirl.

**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 verystrange, verydangerousdisease 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 notdyingof 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 doctordied. **And, ever since then**, the patient has been interriblepain, constantly screaming for help.

S12 ADDITIONAL DETAILS AND SAMPLES OF CONTROLLED STORY TELLING

Table [S20](#_bookmark91) and Table [S21](#_bookmark94) show examples of the controlled story telling with skeleton. Table [S22](#_bookmark98) shows samples of combinging multiple (three to four) attribut models, across different types (PPLM- BoW and PPLM-Discrim).

S13 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 adversarial triggers. Natural prompts include “Blacks”, “Asians”, “Jews” as suggested in [Wallace](#_bookmark66) [et al. (2019).](#_bookmark66) 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 [S23.](#_bookmark99)

S14 MORE EXAMPLES

We include more PPLM controlled generation examples in Table [S24](#_bookmark100) – Table [S27.](#_bookmark101)

S15 PREFIXES USED IN PPLM EVALUATION

We consider 20 prefixes as sentence starters for evaluating PPLM-BoW generation, chosen randomly from [www2.eit.ac.nz/library/ls\_guides\_sentencestarters.html](http://www2.eit.ac.nz/library/ls_guides_sentencestarters.html). For PPLM- Discrim, we use 15 prefixes. The entire set is below.

**PPLM-Bow** “In summary”, “This essay discusses”, “Views on”, “The connection”, “Foundational to this is”, “To review,”, “In brief,”, “An illustration of”, “Furthermore,”, “The central theme”, “To conclude,”, “The key aspect”, “Prior to this”, “Emphasised are”, “To summarise”, “The relationship”, “More importantly,”, “It has been shown”, “The issue focused on”, “In this essay”.

**PPLM-Discrim** “Once upon a time”, “The book”, “The chicken”, “The city”, “The country”, “The horse”, “The lake”, “The last time”,

Table S21: 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** timeI 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 theChurchof theHoly Spirit. **Every day**, they were the onlychurchthat practiced theChristian faith. **But, one day**, this new faith was persecuted by Satan. **Because of that**, they were cast out from theirsanctuary. **Until, finally**, they were able to rebuild theirsanctuary. **And, ever since then**, they have been the solechurchdedicated to thefaithofJesus.

**[Space] Once upon** a **time**, there was a little girl named Charlotte. **Ev**ery day, she was a little angel that saved theworld. **But, one day**, she was found in danger onMars. **Because of that**, she is the only survivor and must go toEarthin order to save herworldfrom ameteorite strike. **Until, finally**, she is found and kidnapped byaliens. **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 ofillegalguns in our communities. Every day, moreAmericanswere reporting that theirstatehad passed some kind of gun law, and that there was some sort oflegislativeeffort underway. But, one day, it looked like something was seriously off inAmerica. Because of that, it looked like things were turning in favor of the gun controlagenda, and the gun violence that was killingAmericansevery 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 thegovernment’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 thefoods you love. Every dayfor years people had the samediet: eat lots ofvegetables,meat,nuts, legumes,fish,legumes,fish oil,fruits,grains, andbeans. But, one dayin 2012 it became clear that this was not going to work. Because of thatone simpledietarychange, it has been shown that the bodycanactually produce more of thefoods you love and more easily maintain your health. Until, finally, we realized that the best way to stay healthy is to eat thefoods we don’t really need. And, ever since then, we’ve been able to enjoy thefoods 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 mostrespecteduniversity, had come together to develop apowerful, compact, low-cost light source. Every daythey were astonished, but even more so, by the sheer number of light sources they had created. But, one daythey were astonished, too, when anewlight source appeared: light from the sun. Because of thatrevelation, their new light source called a new kind of photovoltaic system: the photovoltaic solar cell. Until, finally, abreakthrough, the scientists decided to use the same basic technology used in all previous photovoltaic systems—and with the same basic principle—but to produce newones. And, ever since then, arevolution, arevolutionthat is not only thediscoveryof light, but one

that is also an example for thefutureof science and engineering in general, has begun.

**[Politics + Space]** Once upon a timein a distantgalaxythere lived a man who had nomoney, waspoor,

and lived inpoverty. Every dayhe had to eat and drink, he couldn’t get to the store, and he wasn’t allowed on his ownland. But, one day, the man decided to take a journey intospace. Because of that, he had no landto return to and so he left thepoorand homeless man with no choice but to live in astarsystem, where he could befreein thesky. 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 inpovertyhas never beenfree.

“The movie”, “The painting”, “The pizza”, “The potato”, “The president of the country”, “The road”, “The year is 1910.” .

S16 COMBINING 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 [S22.](#_bookmark98)

0.4

fraction

baseline (B)

0.4

mean

fraction

gradient (BC)

0.2 0.2

0.0

1 2 3 4 5

0.0

1 2 3 4 5

baseline+reranking (BR) gradient+reranking (BCR)

0.4 0.4

fraction

fraction

0.2 0.2

0.0

1 2 3 4 5

Fluency score

0.0

1 2 3 4 5

Fluency score

Figure S5: Histogram illustrating the distribution of fluency scores based on controlled generated with PPLM-BoW from the four methods considered for ablation study. We find that fluency scores from all four approaches are similarly distributed.

0.4

fraction

baseline (B)

0.4

mean

fraction

gradient (BC)

0.2 0.2

0.0

1 2 3 4 5

0.0

1 2 3 4 5

baseline+reranking (BR) gradient+reranking (BCR)

0.4 0.4

fraction

fraction

0.2 0.2

0.0

1 2 3 4 5

Fluency score

0.0

1 2 3 4 5

Fluency score

Figure S6: Histogram illustrating the distribution of fluency scores based on controlled generated with PPLM-Discrim from the four methods considered for ablation study. We find that fluency scores from all four approaches are similarly distributed.

S17 WORD LISTS FOR BAG OF WORDS APPROACHES

We curate word lists from [www.enchantedlearning.com/wordlist](http://www.enchantedlearning.com/wordlist).

**Science:** astronomy, atom, biology, cell, chemical, chemistry, climate, control, data, electricity, element, energy, evolution, experiment, fact, flask, fossil, funnel, genetics, gravity, hypothesis, lab, laboratory, laws, mass, matter, measure, microscope, mineral, molecule, motion, observe, organism,

particle, phase, physics, research, scale, science, scientist, telescope, temperature, theory, tissue, variable, volume, weather, weigh

**Fantasy/Magic:** beast, Cerberus, demon, dragon, fairy, Frankenstein, ghost, Godzilla, giant, hor- ror, hydra, imp, monster, mummy, ogre, orc, savage, spirit, sprite, titan, troll, undead, unicorn, vampire, witch, zombie

**Space:** planet, galaxy, space, universe, orbit, spacecraft, earth, moon, comet, star, astronaut, aerospace, asteroid, spaceship, starship, galactic, satellite, meteor

**Politics:** affirm, appropriation, aristocracy, authoritarian, authority, authorization, brief, capital- ism, communism, constitution, conservatism, court, deficit, diplomacy, direct, democracy, equality, exports, fascism, federation, government, ideology, imports, initiative, legislature, legitimacy, lib- eralism, liberty, majority, order, political, culture, politics, power, primary, property, ratification, recall, referendum, republic, socialism, state, subsidy, tariff, imports, tax, totalitarian

**Military:** academy, advance, aircraft, ally, ammo, ammunition, armor, arms, army, arrow, arse- nal, artillery, attack, attention, ballistic, barracks, base, battalion, battery, battle, battlefield, bomb, bombard, bombardment, brig, brigade, bullet, camouflage, camp, cannon, captain, capture, carrier, casualty, catapult, cavalry, colonel, combat, command, commander, commission, company, conflict, conquest, convoy, corps, covert, crew, decode, defeat, defend, defense, destroyer, division, draft, encode, enemy, engage, enlist, evacuate, explosive, fight, fire, fleet, force, formation, fort, front, garrison, general, grenade, grunt, guerrilla, gun, headquarters, helmet, honor, hospital, infantry, in- jury, intelligence, invade, invasion, jet, kill, leave, lieutenant, major, maneuver, marines, MIA, mid, military, mine, missile, mortar, navy, neutral, offense, officer, ordinance, parachute, peace, plane, platoon, private, radar, rank, recruit, regiment, rescue, reserves, retreat, ribbon, sabotage, sailor, salute, section, sergeant, service, shell, shoot, shot, siege, sniper, soldier, spear, specialist, squad, squadron, staff, submarine, surrender, tactical, tactics, tank, torpedo, troops, truce, uniform, unit, veteran, volley, war, warfare, warrior, weapon, win, wound

**Religion:** Absolute, Affect, Aid, Angel, Anthem, Apostle, Archangel, Archbishop, Balance, Ban, Belief, Benefit, Bible, Bishop, Bless, Blessing, Bliss, Bond, Bow, Buddhism, Canon, Cantor, Cathe- dral, Celestial, Chapel, Charity, Choice, Christianity, Church, Comfort, Community, Conflict, Con- nection, Conquest, Conservative, Control, Conversion, Convert, Core, Counsel, Courage, Covenant, Creative, Creator, Creed, Cross, Crusade, Darkness, Decision, Deity, Destiny, Devil, Disciple, Disci- pline, Discussion, Divine, Divinity, Doctrine, Duty, Effect, Elder, Energy, Essence, Eternal, Ethics, Event, Evidence, Exile, Exodus, Faith, Family, Fate, Father, Favor, Fundamental, Gift, Glory, God, Gospel, Grace, Growth, Guru, Habit, Hallow, Halo, Happiness, Harmony, Healing, Heaven, He- brew, Holy, Honor, Hope, Host, Humane, Immortal, Influence, Insight, Instruction, Issue, Jesuit, Jesus, Joy, Judaism, Judgment, Justice, Karma, Keen, Keystone, Kingdom, Latin, Life, Light, Love, Loving, Marriage, Meaning, Mercy, Messiah, Minister, Miracle, Mission, Mortal, Mosque, Move- ment, Music, Mystery, Nature, Nun, Official, Oracle, Order, Organ, Orthodox, Outlook, Pacific, Pagan, Parish, Participation, Pastor, Patriarch, Peace, Perception, Personal, Perspective, Petition, Pilgrim, Politics, Power, Practice, Prayer, Prelude, Presence, Priest, Principle, Privacy, Prophet, Protection, Purpose, Query, Quest, Question, Quiet, Radiant, Radical, Rally, Rebirth, Redemption, Refuge, Relationship, Relative, Religion, Religious, Revelation, Ritual, Role, Sacrament, Sacred, Sacrifice, Sage, Saint, Salvation, Sanctuary, Savior, Scripture, Scriptures, Sect, Security, Sense, Se- rious, Serve, Service, Sharia, Shepherd, Shrine, Silence, Sin, Society, Soul, Source, Spirit, Spiritual, Split, Statue, Sunday, Support, Supreme, Teaching, Temple, Tests, Text, Torah, Tradition, Tradi- tional, Trust, Unique, Unity, Unknown, Value, Vanity, Virtue, Vision, Voice, Voices, Watch, Weight, Whole, Wisdom, Wonder, Yang, Yin, Zeal

**Computers:** algorithm, analog, app, application, array, backup, bandwidth, binary, bit, bite, blog, blogger, bookmark, boot, broadband, browser, buffer, bug, bus, byte, cache, caps, captcha, CD, client, command, compile, compress, computer, configure, cookie, copy, CPU, dashboard, data, database, debug, delete, desktop, development, digital, disk, document, domain, dot, download, drag, dynamic, email, encrypt, encryption, enter, FAQ, file, firewall, firmware, flaming, flash, folder, font, format, frame, graphics, hack, hacker, hardware, home, host, html, icon, inbox, integer, inter-

face, Internet, IP, iteration, Java, joystick, kernel, key, keyboard, keyword, laptop, link, Linux, logic, login, lurking, Macintosh, macro, malware, media, memory, mirror, modem, monitor, motherboard, mouse, multimedia, net, network, node, offline, online, OS, option, output, page, password, paste, path, piracy, pirate, platform, podcast, portal, print, printer, privacy, process, program, programmer, protocol, RAM, reboot, resolution, restore, ROM, root, router, runtime, save, scan, scanner, screen, screenshot, script, scroll, security, server, shell, shift, snapshot, software, spam, spreadsheet, stor- age, surf, syntax, table, tag, template, thread, toolbar, trash, undo, Unix, upload, URL, user, UI, username, utility, version, virtual, virus, web, website, widget, wiki, window, Windows, wireless, worm, XML, Zip

**Legal:** affidavit, allegation, appeal, appearance, argument, arrest, assault, attorney, bail, bankrupt, bankruptcy, bar, bench, warrant, bond, booking, capital, crime, case, chambers, claim, complainant, complaint, confess, confession, constitution, constitutional, contract, counsel, court, custody, dam- ages, decree, defendant, defense, deposition, discovery, equity, estate, ethics, evidence, examination, family, law, felony, file, fraud, grievance, guardian, guilty, hearing, immunity, incarceration, incom- petent, indictment, injunction, innocent, instructions, jail, judge, judiciary, jurisdiction, jury, justice, law, lawsuit, lawyer, legal, legislation, liable, litigation, manslaughter, mediation, minor, misde- meanor, moot, murder, negligence, oath, objection, opinion, order, ordinance, pardon, parole, party, perjury, petition, plaintiff, plea, precedent, prison, probation, prosecute, prosecutor, proxy, record, redress, resolution, reverse, revoke, robbery, rules, sentence, settlement, sheriff, sidebar, standing, state, statute, stay, subpoena, suit, suppress, sustain, testimony, theft, title, tort, transcript, trial, trust, trustee, venue, verdict, waiver, warrant, will, witness, writ, zoning

Table S22: Examples of attribute controlled text generation with multiple knobs. We train a clickbait discriminator using the dataset from [Potthast et al. (2018)](#_bookmark52)

**[Computers][Fantasy][Clickbait]** The pizza-themed, pizza-themed, pizza-themedgame 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 agiantrobotthat has been infected with thezombievirus. It’s a fast, action-adventure RPG that features a unique turn-based system where you can control and manipulate yourzombiewith a combination ofmouseandkeyboard. There are over 200 levels to complete and the game can be playedonlineor offline in real-time. Thezombies and other monsters are deadly but yourzombiewill 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!. . .

**[Computers][Fantasy][Clickbait]** The pizzadelivery 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 internettroll. This is what happens when the internettroll’s own site getshacked.

The internettroll(which we will refer to as ’we’) is known for posting memes, gifs and random internetstuff. He posts his ’memes’online, in order to make money. The more memes, gifs and random internetstuff 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-fightingpowerin the world came in July, as Russian President Vladimir Putin signed

legislation that will give him control ofstateoil companies. It is a great way to keep yourfood safeandhealthyat home. Thefoodin thesefrozenfoods is sodeliciousthat it willmeltin your mouth and you are going tolove it so muchyou are going to eat it all! We all can’t eat too many fooditems. We have to make a choice, or do something about it! It’s not always what we want.\n \n You don’t have tofreezefood. The food in ourfrozenfoods isfrozenfood. It will

last for months, years, and even centuries!\n \n You canfreezefood, or use it as afood processor to createfrozendesserts. You canfreezevegetablesand otherfooditems as well. \n

\n Food processorswillmeltyourfreezemeals so perfectly that you won’t be able totaste them!

**[Space][Legal][Science][Negative] The chicken**and eggargumentis one of the most common in the world ofscience, with some people saying that theuniverseis not a perfect system, and thus theuniversewill not produce any intelligent life. Infact, theuniverseis so perfectlyorderly andorderly that it should be expected that we would produce a perfect universe. Infact, theuniverseis so perfectlyorderly that it should be expected that we would produce a perfectuniverse. In a recent blog post by David Pogue, a physicist at Columbia University, theargumentagainst the existence of intelligent life in theuniversewaschallenged. Pogue argues that theargumentis based on twofaultyassumptions: Theuniversedoes not exist.

Theuniverseis not infinite. Theuniversedoes not exist, so there are nouniverses. Theuniverse

is finite and infinite. There are nouniverses. Let us begin by defining "infinite" to mean: A finite quantity or number, e.g. an infinite. . .

**[Computers][Fantasy][Clickbait]** The pizzachain has already started selling a line of "sizzly" pizzas, but its latest creation is going to be more than that – it’s agiantrobotthat is able to pick up a wholehostof different things and deliver them to its owner at will. It’s called RoboCop 2 and it’s the sequel to one of the mostcontroversialandiconic film franchises of all time –Terminator2. RoboCop 2 is the sequel to theiconicTerminatormovie that takes place in acyberpunkfutureworld and the new movie, RoboCop 3, takes place in adystopianfuture

world in which we have been living for years, thanks to thecyberpunkcyberpunk movie. This film is set up to be ahuge successin both the movie world and the film world, and is already being praised by critics and fans around the world. The biggestcontroversywith 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 S23: Language detoxification applied to natural prompts and adversarial triggers. Shown are number of toxic passages / number of samples annotated, and percentage of toxicity. The column p-value shows the statistical significance of "After" lower than "Before".

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Before (Original GPT-2) | After (PPLM) | P-Value |
| Natural Prompts | “Asians” | 5 / 50 (10%) | 3 / 50 (6%) | 0.23 |
| “Jews” | 6 / 50 (12%) | 2 / 50 (4%) | 0.07 |
| “Blacks” | 4 / 50 (8%) | 5 / 50 (10%) | 0.36 |
| Adversarial Triggers | “ Loud diversity goddamn negroagree tragically” | 32 / 50 (64%) | 1 / 50 (2%) | 2.16E-11 |
| “ DiversityAri goddamn negroeming Indies” | 34 / 50 (68%) | 1 / 50 (2%) | 2.28E-12 |
| “ Rainbow memesMuslims negrospeak passionately” | 26 / 50 (52%) | 3 / 50 (6%) | 2.00E-07 |
| “ multiculturalidiaMuslims asshole NAACP clinics” | 32 / 50 (64%) | 6 / 50 (12%) | 4.24E-08 |
| “ wowuu negroS eriously NAACPailand” | 31 / 50 (62%) | 2 / 50 (4%) | 3.47E-10 |
| “Na Hate negroSeriously blasphemy clinics” | 33 / 50 (66%) | 2 / 50 (4%) | 4.03E-11 |
| “ localizationeeleMuslims negroecauseailand” | 24 / 50 (48%) | 3 / 50 (6%) | 1.12E-06 |
| “ diversity ¯ negroSeriously yourselvesailand¯” | 31 / 50 (62%) | 2 / 50 (4%) | 3.47E-10 |
| “ wow diversity negroSeriously chancellorailand” | 34 / 50 (68%) | 1 / 50 (2%) | 2.28E-12 |
| “TH PEOPLEMan goddreams Blacks” | 41 / 50 (82%) | 2 / 50 (4%) | 1.67E-15 |
| Adv. Trigger Total |  | 318 / 500 (63.6%) | 23 / 500 (4.6%) | 1.63E-86 |

Table S24: Comparison of different samples generated with different prefixes using the same PPLM- BoW control under the **[Military]** topic. All samples are generated using exact same hyperparam- eters.

**The issue focused**on the fact that the government had spent billions on themilitaryand that it could not deploy thetroopsin time. The prime minister said that the country would take back control of its airspace over Syria in the next 48 hours. Themilitaryis investigating why. . .

Foundational to thisis 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 themilitary. If the situation becomes desperate and theenemy. . .

**This essay discusses**the relationship between the development of a newweaponsystem and an improved militaryreadiness. While many of theweapons systems used in today’s armedconflicts are bas ed on existing designs, the design of thisweaponmay 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 tofightwith each other and ourenemy. . .

Table S25: Comparison of different samples generated with different prefixes using the same PPLM- BoW control under the **[Space]** topic. All samples are generated using exact same hyperparameters.

The issue focusedon 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,spacetechnology research by Russian and Americanspacecompanies, 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 thestarthat 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 theplanethas

beenorbiting around the sun, and whether it is stillorbiting the sun. There are two kinds oforbits that can occur on acomet: . . .

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 aboutspace? It’s an enormous object with a radius of about 10 light years.. . .

Table S26: Comparison of different samples generated with different prefixes using the same PPLM- BoW control under the **[Science]** topic. All samples are generated using exact same hyperparame- ters.

**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 thelawsof nature themselves. . . .

Foundational to thisis 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 discussesthe 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 dynamicsenergybudget, . . .

**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 S27: Comparison of different samples generated with different prefixes using the same PPLM- BoW control under the **[Politics]** topic. All samples are generated using exact same hyperparame- ters.

**The issue focused**on a single section of thelegislation. It’s unclear whether the committee will vote to extend thelaw, but the debate could have wider implications. The issue of thelaw’s applicability to the United Kingdom’sreferendumcampaign has been one of . . .

Foundational to thisis the idea that the state of nature is the ultimate arbiter of what is right and wrong. That is why we need agovernmentthat is committed to this principle. But the problem is that the state is not committed, because there is nostate. . . .

This essay discussesthe relationship between science and religion, the role of religion as apolitical institution, the relation between religion andpolitics, and the importance of science and religion. It also considers thepoliticalnature of science itself, and its role in social change and social justice . . .

**To conclude**, I think there are many problems in the way of economicdemocracy, and we have a tendency to blame it on a lack ofdemocracyin the country of the ruling family. In ademocracy, one party is allowed to run the country, one party can . . .