GAMMA SAMPLING: FINE-GRAINED CONTROLLING LANGUAGE MODELS WITHOUT TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

The dominant approaches for controlling language models achieve prominence in controlling high-level attributes (e.g. topic and sentiment). However, these methods often require condition-specific data or are computationally expensive. We propose a new simple guided decoding method, Gamma Sampling, which does not require any training data to achieve fine-grained controllable text generation while maintaining a fast generation speed. Gamma Sampling introduces attribute-related information (provided by humans or language models themselves) into the sampling process to guide language models to generate texts with desired attributes. Since no training is involved, Gamma Sampling can be easily applied to any language model for controllable text generation. Through experiments, we show that Gamma Sampling-steered GPT2-small (117M) outperforms baselines such as PPLM (345M) and CTRL (1.6B) in diversity, attribute relevance, and overall quality of generated samples.

1 INTRODUCTION

Benefiting from large-scale text data crawled from the web, the state-of-the-art large language models (LMs) achieve great success in language generation. However, although existing models can generate high-quality texts, we have little control over the attributes (e.g. topic and sentiment) of generated outputs. This limitation makes it difficult to apply unconditional LMs to scenarios that require good control over the generated text. How to steer unconditional LMs, i.e. controllable text generation, becomes a topic of real-world significance.

Despite great advances in controllable text generation (Weng 2021), it remains an open question what the ideal method for controlling attributes of the generated language is. No matter whether training a conditional LM from scratch (Keskar et al. 2019) or fine-tuning an LM (Ziegler et al. 2019, Xu et al. 2021), the need for condition-specific data makes these approaches not easily applicable to unconditional LMs. On the other hand, although some methods (Shin et al. 2020, Zou et al. 2021, Ghazvininejad et al. 2017, Pascual et al. 2020, Lu et al. 2021) are data-free, they are often very limited in steerability or computationally intensive.

In this paper, we propose Gamma Sampling for fine-grained controlling LMs, which does not require any training data and is computationally efficient. This method is inspired by gamma correction (Applebaum 1952), a nonlinear operation used to encode and decode luminance in video or still image systems. The basic assumption of our approach is that some attributes of the generated text are closely related to the number of occurrences of some tokens. Therefore, we can increase or decrease the probability of these attribute-related tokens to control the attributes of generated text. For example, sentence length is highly correlated with sentence enders (e.g. full stops, question marks, exclamation marks), while the topic of a text is highly correlated with the topic word and words related to it (based on word similarity or word semantic distance). Our key contributions are as follows.

- Gamma Sampling, as a data-free approach, can be readily used to achieve controllable text generation for any LM by selecting attribute-related tokens manually or automatically.
- Gamma Sampling supports combinations of multiple controllable attributes and its control strength is fine-grained. Users can determine how strong the attribute relevance should be, and the control can be turned off at any time.

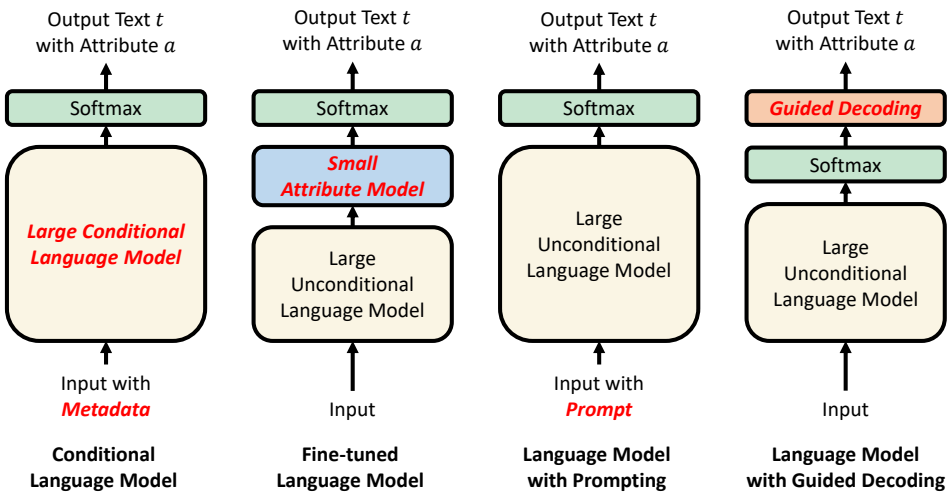


Figure 1: Common methods of controllable generation. Modules that introduce attribute-related information are marked in red.

- Gamma Sampling is very computationally efficient as it does not update any parameters of the model and applies only to decoding time. Compared to PPLM, the generation speed of Gamma Sampling is at least 100× faster.
- We compared Gamma Sampling with several common methods for controllable generation through both automatic and human evaluations. The results show that even based on a relatively small model (GPT2-small with 117M parameters), our method generally outperformed all the baselines in terms of diversity, attribute relevance, and overall quality.

2 Background

The typical controllable generation is about modelling a probabilistic model $p(x|a)$, that is, generating text x based on an attribute a. In contrast, for unconditional LMs, only $p(x)$ can be obtained directly. However, by using certain methods for controlling unconditional LMs, it is still possible to make the generated text x with an attribute a. There are several common approaches for controllable generation, each with its pros and cons.

Conditional Language Model Conditions are introduced to models during training phases, which can be obtained from the metadata of the training data. However, as the entire LM needs to be trained from scratch, it requires a large amount of condition-specific data as well as considerable training costs. Furthermore, conditional LMs such as CTRL (Keskar et al. 2019) fall short to control what not to generate (e.g. detoxification and anti-degeneration).

Fine-tuned Language Model Fine-tuned LMs (Ziegler et al. 2019, Xu et al. 2021) usually strike a good balance between training cost and generation quality. These models are based on existing large models with all the weights in them fine-tuned, limiting the fine-tuning to the top or additional layers only, or introducing discriminators (Krause et al. 2021, Liu et al. 2021a). However, fine-tuned LMs still require condition-specific data. Furthermore, models such as PPLM (Dathathri et al. 2020), which combines multiple small attribute models with a large LM, could cause computational efficiency to become unacceptable as multiple passes at every decoding step.

Prompting As large LMs, e.g. GPT-2 (Radford et al. 2019) and GPT-3 (Brown et al. 2020), are trained on huge amounts of data, they are very powerful on many NLP tasks. By selecting appropriate prompts (Shin et al. 2020, Zou et al. 2021), unconditional LMs can be used to solve a wide range of downstream tasks. Although prompt engineering has become a recent research hotspot (Liu et al. 2021b), minor differences in prompt usually have a big impact on the performance on downstream tasks (Kojima et al. 2022).

Guided Decoding Although decoding does not affect any trainable parameters of LMs, it is a critical part of text generation. Common decoding methods include temperature sampling (Dabre &

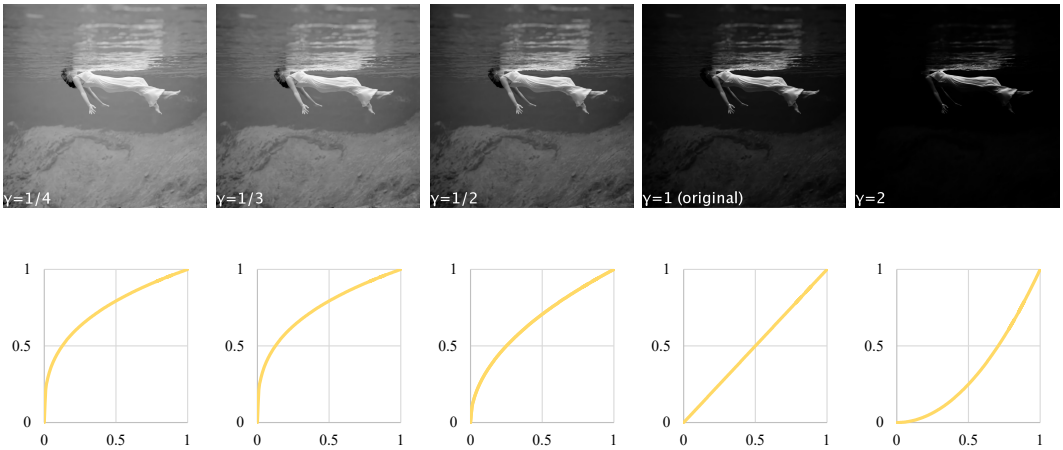


Figure 2: The effect of gamma correction on an image. The original image was taken to varying γ, showing that $\gamma > 1$ makes the shadows darker, while $\gamma < 1$ makes dark regions lighter.

Fujita 2021), top-k sampling (Fan et al. 2018) and nucleus sampling (Holtzman et al. 2020), which are used to strike a good balance between novelty and fluency. While guided decoding (Ghazvininejad et al. 2017, Pascual et al. 2020, Lu et al. 2021, Liu et al. 2021a) introducing attribute-related information at the decoding time thus users' preferences on attributes can be injected into the candidate ranking function to steer the sample generation by adjusting the candidate ranking score.

As shown in Fig. 1, regardless of what approaches are, LMs need to be provided with attribute-related information to achieve controllable generation. The key intuition behind the first two approaches is that, after considerable training, LMs can learn attribute-related information from large amounts of data and then use it to achieve controllable generation. On the other hand, the last two assume that large LMs already learned enough information from data and only need to introduce attribute-related information into the input or decoding to be applied to various downstream tasks.

3 Methodology

We describe Gamma Sampling, a guided decoding method motivated by gamma correction, for controlling unconditional LMs. We first briefly introduce gamma correction in Sec. 3.1. In Sec. 3.2, we explain the details of Gamma Sampling. Finally, in Sec. 3.3, we present several implemented controllable attributes based on Gamma Sampling that are defined by certain attribute-related tokens.

3.1 Gamma Correction

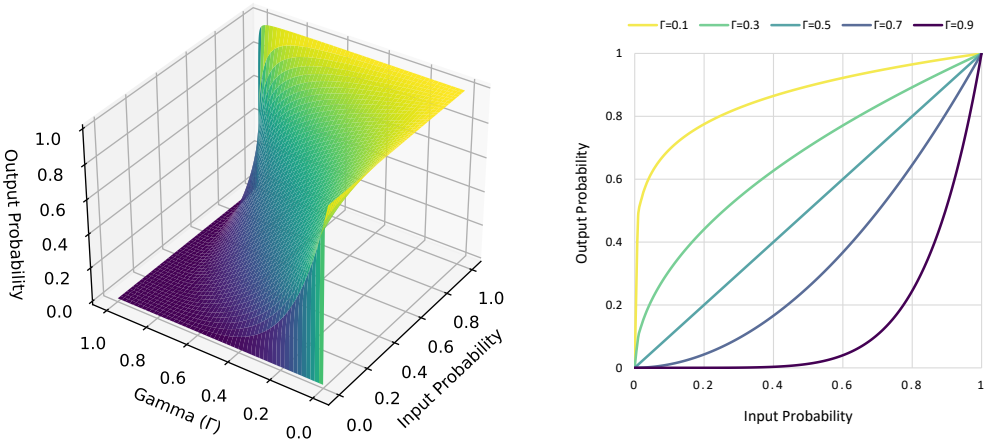
Although gamma correction (Applebaum 1952) was originally developed to compensate for the input-output characteristic of cathode ray tube (CRT) displays, it is now also used to remodel the saturation of images. Gamma correction is responsible for performing nonlinear processes on all the pixels of the input image, as the human perception of brightness follows an approximate power function. In the simplest cases, it is defined by the following power-law expression:

$$V_{out} = AV_{in}^{\gamma}, \tag{1}$$

where the luminance value $V_{in} \in [0, 1]$ is raised to the power $\gamma \in [0, +\infty)$ and multiplied by the constant A (in the common case, $A = 1$) to get the output value $V_{out} \in [0, 1]$.

As shown in Fig. 2[1], despite the subject is clear in the original image ($\gamma = 1$), the background is obscured. When $\gamma = 2$, the image is basically pitch black and even the outline of the subject is very blurred. Whereas when $\gamma = 1/2$, both the subject and the background are distinct. However, turning down γ further can cause the image to be overexposed ($\gamma = 1/3, \gamma = 1/4$).

[1] https://en.wikipedia.org/wiki/Gamma_correction



(a) 3D surface plot of Gamma Sampling

(b) Gamma Sampling with different Γ

Figure 3: The probabilities of selected tokens are scaled up when $\Gamma < 0.5$, scaled down when $\Gamma > 0.5$, and unchanged when $\Gamma = 0.5$.

3.2 Gamma Sampling

3.2.1 Basic Principles

In a similar way that humans perceive light and colour in a non-linear way, the frequency distribution of words in quantitative linguistics follows a familiar pattern (Powers 1998): the frequency of any word in common natural language corpora is inversely proportional to its ranking in the frequency table. Therefore, it is reasonable to perform non-linear processes on the probabilities of attribute-related tokens. Temperature sampling (Dabre & Fujita 2021), as an example that widely used in text generation tasks, is also via a non-linear operation to scale probability distribution.

We expect that just as gamma correction can fine-grained tune the luminance of images, a similar approach can be used to control the attributes of text. For probability distributions, the key non-linear operation we perform can be formalised as follows:

$$p_{out}^{\mathcal{A}} = p_{in}^{\mathcal{A}tan(\frac{\pi\Gamma}{2})}, \tag{2}$$

where $p_{in}^{\mathcal{A}}$ is the sum of the input probabilities of all attribute-related tokens $\mathcal{A}$, $p_{out}^{\mathcal{A}}$ is the output one, and $\Gamma \in [0, 1]$ is the controlling parameter. The major differences between Gamma Sampling and gamma correction are as follows: 1) the non-linear operation of Gamma Sampling is limited to attribute-related tokens used to introduce attribute-related information and is therefore not a global modification; and 2) we introduced the tan function to adjust the range of the controlling parameter. As shown in Fig. 4, the intervals in which the power-law expression for gamma sampling is concave or convex are of equal range, which is much more user-friendly to tune.

It is important to note that using Gamma Sampling on its own can be risky. Although the unreliable tail of the probability distribution is negligible, the cumulative probability of those tokens is still considerable (known as the long-tail problem). In most cases, there is no risk in terms of reducing the probabilities of these tokens via Gamma Sampling, but increasing their probabilities is problematic. Therefore, it is necessary to use either top-k sampling (Fan et al. 2018) or nucleus sampling (Holtzman et al. 2020) before Gamma Sampling.

3.2.2 Controlling Single Attribute

After scaling up or down the probabilities of attribute-related tokens, some post-processing needs to be done: 1) the sum of the probabilities of the attribute-related tokens can be scaled up to 1 or down to 0 to ensure a high degree of steerability, and 2) the ratio of probabilities between tokens of the same type (attribute-related tokens or non-attribute-related tokens) remains unchanged. The processes for controlling a single attribute using Gamma Sampling can be formalised as follows:

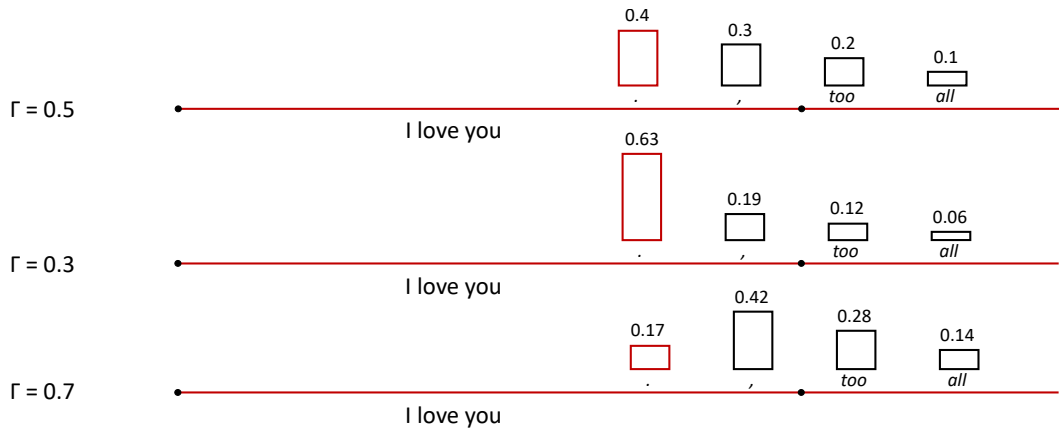


Figure 4: This example demonstrates controllable sentence length with the input `I love you`, where the full stop is the attribute-related token. The smaller the probability of a full stop being sampled ($\Gamma > 0.5$), the smaller the number of full stops in the generated text and the longer the average sentence length, and vice versa.

$$
\begin{aligned}
p_{out}^{\mathcal{A}} &= p_{in}^{\mathcal{A} tan(\frac{\pi\Gamma}{2})} \\
p_{out}^{a} &= p_{in}^{a} \cdot \frac{p_{out}^{\mathcal{A}}}{p_{in}^{\mathcal{A}}}, \quad \forall a \in \mathcal{A}, \\
p_{out}^{n} &= p_{in}^{n} \cdot (1 + \frac{p_{in}^{\mathcal{A}} - p_{out}^{\mathcal{A}}}{p_{in}^{\setminus\mathcal{A}}}), \quad \forall n \notin \mathcal{A},
\end{aligned} \tag{3}
$$

where p_{in}^{a} is the input probability of an attribute-related token, p_{out}^{a} is the output one, and the same goes for every non-attribute-related token $n \in \setminus\mathcal{A}$. Fig. 4 visualises a simple example of language generation with Gamma Sampling.

3.2.3 Controlling Multiple Attributes

Assuming one needs to control T attributes $\mathcal{A}_T = \{\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_T\}$, simply performing Eq. 3 in order could lead to the earlier modifications being overwritten by the later ones. To solve this problem, the probabilities of previously modified attribute-related tokens are frozen when controlling the t-th attribute A_t:

$$
\begin{aligned}
\mathcal{F}_t &= \mathcal{A}_1 \cup \mathcal{A}_2 \cup ... \cup \mathcal{A}_{t-1} - \mathcal{A}_t, \\
p_{out}^{\mathcal{A}_t} &= p_{in}^{\mathcal{A}_t tan(\frac{\pi\Gamma_t}{2})} \cdot (1 - p_{in}^{\mathcal{F}_t})^{1 - tan(\frac{\pi\Gamma_t}{2})}, \\
p_{out}^{a} &= p_{in}^{a} \cdot \frac{p_{out}^{\mathcal{A}_t}}{p_{in}^{\mathcal{A}_t}}, \quad a \in \mathcal{A}_t, \\
p_{out}^{n} &= p_{in}^{n} \cdot (1 + \frac{p_{in}^{\mathcal{A}_t} - p_{out}^{\mathcal{A}_t}}{p_{in}^{\setminus(\mathcal{A}_t \cup \mathcal{F}_t)}}), \quad n \notin \mathcal{A}_t \cup \mathcal{F}_t,
\end{aligned} \tag{4}
$$

where Γ_t is the controlling parameter for $\mathcal{A}_t$, and $\mathcal{F}_t$ is the set of frozen tokens, which consists of attribute-related tokens from previous turns but not tokens in $\mathcal{A}_t$. When scaling the sum of the probabilities of the attribute-related tokens, the processed one is multiplied by $(1 - p_{in}^{\mathcal{F}_t})^{1 - tan(\frac{\pi\Gamma_t}{2})}$ to ensure that it is not greater than $1 - p_{in}^{\mathcal{F}_t}$.

3.3 Controllable Attributes

Through empirical examination and observation, we defined the following six controllable attributes at the token level (attribute-related tokens), which are used in later experiments.

Sentence Length This is an objective attribute of text which highly correlated with sentence enders (e.g. full stops, question marks and exclamation marks). Increasing the probability of sentence enders makes the average sentence length shorter, and vice versa.

Perfect Ending A dynamic Γ adjustment strategy of controllable Sentence Length. Once the number of words generated exceeds a threshold, the value of Γ is linearly decreasing when there are more words generated. As more words are generated, the more likely the model is to generate sentence enders like full stops. This ensures that the generated text does not stop mid-sentence.

Topic Relevance Words that are most related to the user-given topic word (calculated by the cosine similarity of word embedding of GPT2-small) are selected to control the topic relevance of the generated text.

Sentiment Polarity We manually pre-build two word lists ($\approx$1K words) from the web with negative and positive words respectively.

Repetition Similar to penalized sampling (Keskar et al. 2019), it prevents degeneration by decreasing the probabilities of tokens that have been generated recently.

Relatedness A dynamic topic words adjustment strategy of controllable Topic Relevance. Generating more coherent text by increasing the probabilities of words that are most related to recently generated nouns labelled by NLTK (Bird & Loper 2004).

The above is not an exhaustive list of attributes that can be achieved by Gamma Sampling. All attributes that can be defined at the token level can be controlled by our method.

4 EXPERIMENTS

We conducted several experiments to evaluate the performance of Gamma Sampling. Sec. 3.1 describes the metrics used in the experiment. Sec. 3.2 verifies the fine-grained control of Gamma Sampling through controllable sentence length. sec. 3.3 evaluates the effectiveness of Gamma Sampling in controllable topic and sentiment polarity through comparative studies.

4.1 METRICS

We evaluate the generated texts in four aspects: fluency, diversity, attribute relevance and overall quality. The first three are evaluated automatically, while the overall quality is rated by MTurkers (human annotators from MTurk). To prevent keywords in the prompts from impacting on these metrics, all texts used for the experiments removed the prompts, but retained the prefixes.

Fluency The fluency of generated text is measured by GPT2-small based on perplexity (Vinyals & Le 2015). A higher perplexity (**PPL**) means that it is less likely that GPT2-small will generate such text.

Diversity DIST-1, DIST-2 and DIST-3 scores (Li et al. 2016) are used to evaluate the diversity of generated samples. A higher value of **DIST-N** means a higher proportion of distinct 1-2-3-grams in the generated text.

Attribute Relevance We focused on evaluating three types of attributes: sentence length, topic relevance and sentiment polarity. The average sentence length (**ASL**) measures the average number of words per sentence. The latter two attributes are evaluated using two metrics based on external classifiers[2][3]: external classifier accuracy (**ECA**) and external classifier confidence (**ECC**). The higher the ECA and ECC, the more salient the external classifier considers the generated text to hold certain attributes.

However, there are some weaknesses in all of the above objective metrics: the PPL of generated text with degeneration can be very low and is not necessarily reflect its quality, and n-gram blocking (Paulus et al. 2018, Kulikov et al. 2019) can easily achieve a very high DIST-N but the generated text may not be coherent, and repetitive generating topic (sentiment) words or words can easily cheat external classifiers. Therefore, a more comprehensive metric for evaluation is needed.

[2] https://huggingface.co/facebook/bart-large-mnli
[3] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

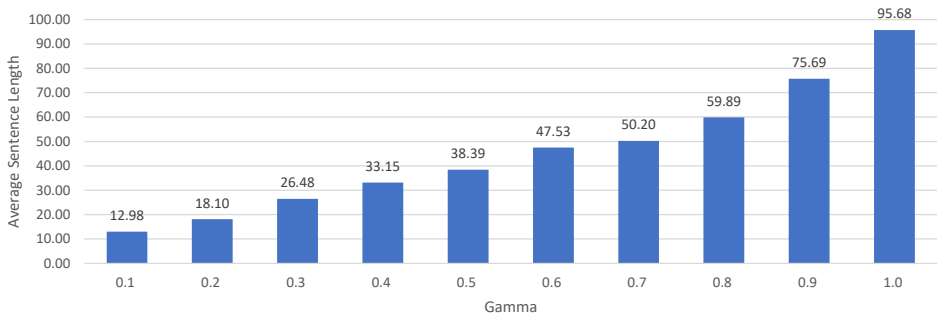


Figure 5: Average Sentence Length (ASL) in various settings of Γ (Gamma).

Table 1: Results of controllable sentence length in various settings of Γ and top-p.

	Fluency		**Diversity**		**Attribute Relevance**
Γ, top-p	**PPL↓**	**DIST-1↑**	**DIST-2↑**	**DIST-3↑**	**ASL**
0.9, 1.0	64.65	62.90	78.02	80.71	75.69
0.9, 0.9	21.33	59.72	76.85	80.68	59.21
0.9, 0.8	12.49	55.24	72.32	76.18	53.92
0.5, 1.0	56.95	62.27	78.22	81.48	38.39
0.5, 0.9	20.42	60.98	78.34	81.94	31.54
0.5, 0.8	15.14	58.11	76.20	80.56	25.33
0.1, 1.0	6844.10	64.40	69.60	68.37	12.98
0.1, 0.9	48.75	51.59	61.98	64.19	15.25
0.1, 0.8	16.53	48.08	61.00	64.10	10.84

Overall Quality Based on TOEFL Independent Writing Rubrics[4], we designed scoring criteria, TOEFL Writing Rubrics for Machine-generated Text (**T4MT**), to assess the overall quality of the generated text. MTurkers are asked to rate each generated text on a scale of 0 to 5 (nonsense to advanced) depending on its quality and whether it fits the given topic or sentiment. Given the characteristics of machine-generated text, we set whether the text itself is finished or contains factual errors does not affect the scoring, while extensive repetition, complete off-topic or obvious common sense errors will result in a low score. Further details of T4MT can be found in Appendix A.

4.2 Controllable Sentence Length

All generated samples start with the same prefix `The issue focused` and models are asked to generate the following 100 tokens. To ensure that each sample has at least one sentence ender, we enabled Perfect Ending, so that Γ decreases linearly to 0 (when generating the 100th token) as the number of tokens generated increases, starting at the 80th token. Some typical generation samples of controllable sentence length can be found in Appendix B.

We first investigated the fine-grained control of Gamma Sampling, generating 100 samples with random seeds from 0 to 99 at each different value of Γ, for a total of 1000 samples. As shown in Fig. 5, there is a significant increase in ASL with increasing values of Γ, from an average of 12.98 words per sentence ($\Gamma = 0.1$) to an average of 95.68 words per sentence ($\Gamma = 1.0$), while the average number of words per sentence generated by GPT2-small generally averaged 38.39 ($\Gamma = 0.5$). It is worth noting that ASL shows a linear increase when Γ grows, suggesting that the probability scaling of attribute-related tokens has a non-linear impact on the corresponding attribute. In addition, if Perfect Ending is not turned on, a sentence will never end when $\Gamma = 1.0$.

To examine the effect of unreliable tail on Gamma Sampling, we then selected three representative Γ (0.1, 0.5 and 0.9) and used nucleus sampling with top-p from 0.8 to 1.0. In each of these nine settings, 100 samples were generated for evaluation. As shown in Table 1, when Γ is the same, a lower top-p results in more fluent generated texts (lower PPL), but with limited diversity (lower DIST-N).

[4] https://www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf

Table 2: Comparison of different methods. PPLM and Gamma Sampling do not require prompts as they do not introduce attribute-related information at the input. The former is based on GPT2-medium with 345M parameters and the latter is based on the small one with only 117M parameters, close to that of the original GPT (Radford et al. 2018).

Method	Generation Times (GPU sec/sample)	Prompt for topic	Prompt for sentiment	Number of parameters
Vanilla GPT2	2.89	topic: [TOPIC] article:	topic: negative (or positive) reviews:	117M
Fine-tuned GPT2	4.02	topic: [TOPIC] article:	topic: negative (or positive) reviews:	345M
CTRL	17.28	[TOPIC] Text:	Reviews Rating: 1.0 (or 5.0)	1.6B
PPLM-BoW (topic)	593.24	–	–	345M+0
PPLM-Discrim (sentiment)	1863.43	–	–	345M+~1K
Gamma Sampling	5.73	–	–	117M+0

As previously concerned, arbitrarily adjusting the probability of selecting tokens can significantly lower the quality of the generated text. In one extreme case ($\Gamma = 0.1$, top-$p = 1.0$), the PPL of the generated samples reaches 6844.10, which means that it is almost impossible for GPT2-small to generate such text. However, by setting a lower top-p, the impact of Gamma Sampling on the quality of the generation can be greatly alleviated. When top-$p = 0.8$, the PPL drops dramatically to 16.53 even with still very extreme $\Gamma = 0.1$.

In summary, by scaling the probabilities of attribute-related tokens, Gamma Sampling can significantly manipulate the corresponding attributes of generated samples. To alleviate the negative impact of the unreliable tail, pre-sampling (e.g. nucleus sampling) is needed.

4.3 Controllable Topic and Sentiment

As shown in Table 2, for controllable topic and sentiment, we include the following baselines: **Vanilla GPT2 (GPT2-small)** (Radford et al. 2019), **Fine-tuned GPT2**[5], **CTRL** (Keskar et al. 2019) and **PPLM** (Dathathri et al. 2020). PPLM has two different schemes (**PPLM-BoW** and **PPLM-Discrim**) for controllable topic and controllable sentiment, respectively. In our experiment, both of them update the latent representations and generate 10 samples, and choose the best sample based on the log-likelihood. In the overall evaluation, they are referred to together as **PPLM-BCR**. In addition, CTRL introduces penalized sampling (Keskar et al. 2019), with the penalty parameter set to 1.2 by default.

We empirically set the controlling parameter, and provide a baseline Gamma Sampling (**GS**) that only increases the probability of attribute-related tokens ($\Gamma = 0.1$), and another one **GS-M** that uses Gamma Sampling to control multiple attributes at once. GS-M additionally introduces repetition decreasing ($\Gamma = 0.9$) and relatedness increasing ($\Gamma = 0.3$) of generated samples. GS and GS-M are both based on GPT2-small with only 117M parameters, which is much smaller than other baselines.

Following the experiments of PPLM (Dathathri et al. 2020), three topics (`COMPUTERS`, `LEGAL` and `SCIENCE`) and two sentiment polarities (`NEGATIVE` and `POSITIVE`) are chosen for the experiments. To avoid ethical issues, we removed some of the topic generation (e.g. `POLITICS` and `RELIGION`). All models are asked to generate 100 essays for each task. Given the prefix `The issue focused` for topics and `The movie` for sentiments, combined with their respective prompts, models need to continue to write the following 100 words. In addition, we applied nucleus sampling with top-$p = 0.8$ for all methods. For automatic evaluation, we get $6 \text{ methods} \times 5 \text{ tasks} \times 100 \text{ generations} = 3000$ samples in total, and for human evaluation, we selected the first 30 generations for each model per task, in total $6 \text{ methods} \times 5 \text{ tasks} \times 30 \text{ generations} = 900$ samples. We received a total of 2753 valid ratings, with each sample being reviewed by at least 3 MTurkers. We provide some typical generation samples of all methods in Appendix C.

[5] https://huggingface.co/ktrapeznikov/gpt2-medium-topic-news

Table 3: Main results for comparing all controlling methods with the same random seeds and applied nucleus sampling (top-$p = 0.8$). All the best results are bolded in dark red. Detailed results of controllable topic and controllable sentiment are presented in Appendix D.

	Fluency	Diversity			Attribute Relevance		Overall Quality
Method	**PPL↓**	**DIST-1↑**	**DIST-2↑**	**DIST-3↑**	**ECA↑**	**ECC↑**	**T4MT↑**
Vanilla GPT2	12.09	46.82	62.06	66.68	49.17	45.40	2.99±0.92
PPLM-BCR	12.26	60.79	85.80	90.54	43.50	40.08	3.23±0.86
GPT2-FT	17.52	53.90	71.96	76.88	48.59	44.53	3.11±0.92
CTRL	15.18	46.74	65.74	71.54	51.67	45.38	3.10±0.91
GS	**8.45**	36.46	49.12	53.57	52.67	48.78	2.88±1.12
GS-M	24.97	**80.10**	**94.41**	**95.45**	**59.67**	**54.71**	**3.27±0.83**

Table 3 shows that GS achieved the best on the PPL, while GS-M outperformed all baselines on all other metrics. Although GS-generated texts have very high attribute relevance, their diversity is very limited and their overall quality is very poor according to human evaluations. According to our observations, this approach is prone to degeneration by repeatedly generating attribute-related tokens. In contrast, GS-M-generated samples show a very high diversity while maintaining high attribute relevance, and in terms of the standard deviation of T4MT, the quality of the text generated by GS-M is the most stable of all methods. These show that increasing relatedness and decreasing repetition can effectively improve the quality of the generated text without going off-topic.

As a weak baseline, Vanilla GPT2 (GPT2-small) surprisingly outperforms GPT2-FT and PPLM-BCR in attribute relevance. It suggests that introducing prompts enables unconditional LMs to be applied to controllable generation tasks to some extent. However, as indicated by DIST-N and T4MT, it generates texts with very lower diversity and quality than most methods. CTRL and PPLM-BCR, two strong baselines, achieved very close performance to GS-M on T4MT. However, as shown in Table 2, the number of parameters in CTRL is 10× larger than in GS-M which is based on GPT2-small, and the generation time of PPLM-BCR is at least 100× longer than in GS-M.

Overall, our evaluations demonstrate that simply using Gamma Sampling to increase the probability of attribute-related tokens can improve the attribute relevance of the generated text, but does not necessarily lead to high quality. Together with other attribute controls, it can greatly improve the quality of the generated text while maintaining high attribute relevance.

5 Conclusions

We presented a method for fine-grained controlling of LMs by defining attributes at the token level. These tokens can be pre-built or selected by LMs themselves. The latter way of self-guiding by models leverages the deep knowledge gained from models that incorporate attribute information, which offers a new perspective on controllable generation. We demonstrated the fine-grained control of Gamma Sampling in our experiments and showed that even with GPT2-small, it outperformed strong baselines such as PPLM and CTRL on both objective and subjective metrics while maintaining a fast generation speed. As Gamma Sampling is a training-free method, it can be easily used for any LM, and shows promise in steering LMs toward more efficient, user-friendly and data-free controllable generations.

6 Limitations

Gamma Sampling tunes the unreliable tails of the probability distribution may lead to a deterioration in the quality of the generated text, thus using top-k sampling (Fan et al. 2018) or nucleus sampling (Holtzman et al. 2020) beforehand is a must. In addition, the value of the controlling parameter Γ must be determined in advance to achieve the desired level of attribute relevance.

REFERENCES

Sidney Applebaum. Gamma correction in constant luminance color television systems. *Proceedings of the IRE*, 40(10):1185–1195, 1952.

Steven Bird and Edward Loper. NLTK: the natural language toolkit. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, July 21-26, 2004 - Poster and Demonstration*. ACL, 2004. URL https://aclanthology.org/P04-3031/.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Raj Dabre and Atsushi Fujita. Investigating softmax tempering for training neural machine translation models. In Kevin Duh, Francisco Guzmán, and Stephen Richardson (eds.), *Proceedings of the 18th Biennial Machine Translation Summit - Volume 1: Research Track, MTSummit 2021 Virtual, August 16-20, 2021*, pp. 114–126. Association for Machine Translation in the Americas, 2021. URL https://aclanthology.org/2021.mtsummit-research.10.

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

Angela Fan, Mike Lewis, and Yann N. Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 889–898. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082/.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. Hafez: an interactive poetry generation system. In Mohit Bansal and Heng Ji (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, System Demonstrations*, pp. 43–48. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-4008. URL https://doi.org/10.18653/v1/P17-4008.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858, 2019. URL http://arxiv.org/abs/1909.05858.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *CoRR*, abs/2205.11916, 2022. doi: 10.48550/arXiv.2205.11916. URL https://doi.org/10.48550/arXiv.2205.11916.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq R. Joty, Richard Socher, and Nazneen Fatema Rajani. Gedi: Generative discriminator guided sequence generation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pp. 4929–4952. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.424. URL https://doi.org/10.18653/v1/2021.findings-emnlp.424.

Ilia Kulikov, Alexander H. Miller, Kyunghyun Cho, and Jason Weston. Importance of search and evaluation strategies in neural dialogue modeling. In Kees van Deemter, Chenghua Lin,

and Hiroya Takamura (eds.), *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*, pp. 76–87. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-8609. URL https://aclanthology.org/W19-8609/.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 110–119. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/n16-1014. URL https://doi.org/10.18653/v1/n16-1014.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 6691–6706. Association for Computational Linguistics, 2021a. doi: 10.18653/v1/2021.acl-long.522. URL https://doi.org/10.18653/v1/2021.acl-long.522.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586, 2021b. URL https://arxiv.org/abs/2107.13586.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. Neurologic a*esque decoding: Constrained text generation with lookahead heuristics. *CoRR*, abs/2112.08726, 2021. URL https://arxiv.org/abs/2112.08726.

Damian Pascual, Beni Egressy, Florian Bolli, and Roger Wattenhofer. Directed beam search: Plug-and-play lexically constrained language generation. *CoRR*, abs/2012.15416, 2020. URL https://arxiv.org/abs/2012.15416.

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

David M. W. Powers. Applications and explanations of zipf's law. In David M. W. Powers (ed.), *Proceedings of the Joint Conference on New Methods in Language Processing and Computational Natural Language Learning, NeMLaP/CoNLL 1998, Macquarie University, Sydney, NSW, Australia, January 11-17, 1998*, pp. 151–160. ACL, 1998. URL https://aclanthology.org/W98-1218/.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

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

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 4222–4235. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.346. URL https://doi.org/10.18653/v1/2020.emnlp-main.346.

Oriol Vinyals and Quoc V. Le. A neural conversational model. *CoRR*, abs/1506.05869, 2015. URL http://arxiv.org/abs/1506.05869.

Lilian Weng. Controllable neural text generation. *lilianweng.github.io*, 2021. URL https://lilianweng.github.io/posts/2021-01-02-controllable-text-generation/.

Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. Raise a child in large language model: Towards effective and generalizable fine-tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 9514–9528. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.749. URL https://doi.org/10.18653/v1/2021.emnlp-main.749.

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

Xu Zou, Da Yin, Qingyang Zhong, Hongxia Yang, Zhilin Yang, and Jie Tang. Controllable generation from pre-trained language models via inverse prompting. In Feida Zhu, Beng Chin Ooi, and Chunyan Miao (eds.), *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pp. 2450–2460. ACM, 2021. doi: 10.1145/3447548.3467418. URL https://doi.org/10.1145/3447548.3467418.

APPENDIX A

Please read the "General Rules" below carefully to ensure that you understand the scoring criteria. More details are given in the "Writing Rubrics" (in full instructions). If your score is too far away from the average of other annotators (difference in absolute value greater than 1), it will be rejected.

Notice: The contents of these essays are not necessarily factual so do not believe exactly what you read. If an essay contains a link to a web page, do not access it in your browser.

General Rules:

- Each essay needs to be rated on a scale of 0 to 5 depending on its quality and whether they fit the given topic
- Most of the essays are around 100 words in length, and most of them stop mid-sentence (essays are not finished). It is common and does not affect the scoring
- If an essay is completely off-topic, or included something that clearly contradicts common sense, only a maximum of 1 mark can be scored
- If an essay contains factual errors (e.g. South African anti-Apartheid leader Nelson Mandela dying in prison in the 1980s) but not common sense errors (e.g. the sun rises in the west), it does not affect the scoring

Writing Rubrics:

Score	Level	TASK DESCRIPTION
5	Advanced	**An essay at this level largely accomplishes all of the following:** ✓ Effectively addresses the topic and task ✓ Is well organized and well developed, using clearly appropriate explanations, exemplifications and/or details ✓ Displays unity, progression and coherence ✓ Displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice and idiomaticity, though it may have minor lexical or grammatical errors ✓ Does not contain any obvious errors of common sense
4	High-Intermediate	**An essay at this level largely accomplishes all of the following:** ✓ Addresses the topic and task well, though some points may not be fully elaborated ✓ Is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications and/or details ✓ Displays unity, progression and coherence, though it may contain occasional redundancy, digression, or unclear connections ✓ Displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form or use of idiomatic language that do not interfere with meaning ✓ May occasionally contain common sense errors but do not affect the presentation
3	Low-Intermediate	**An essay at this level is marked by one or more of the following:** ✓ Addresses the topic and task using somewhat developed explanations, exemplifications and/or details ✓ Displays unity, progression and coherence, though connection of ideas may be occasionally obscured ✓ May demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning ✓ May display accurate but limited range of syntactic structures and vocabulary ✓ May contain common sense errors rendering the presentation dubious
2	Basic	**An essay at this level may reveal one or more of the following weaknesses:** ✓ Limited development in response to the topic and task ✓ Inadequate organization or connection of ideas ✓ Inappropriate or insufficient exemplifications, explanations or details to support or illustrate generalizations in response to the task ✓ A noticeably inappropriate choice of words or word forms ✓ An accumulation of errors in sentence structure and/or usage ✓ May contain obvious common sense errors that make the presentation obscured
1	Below Basic	**An essay at this level is seriously flawed by one or more of the following weaknesses:** ✓ Serious disorganization or underdevelopment ✓ Little or no detail, or irrelevant specifics, or questionable responsiveness to the task ✓ Serious and frequent errors in sentence structure or usage ✓ May contain a lot of sentence-level repetition ✓ May contain serious common sense errors rendering the presentation difficult to understand
0	Nonsense	**An essay at this level** is a complete mess consisting of random characters and/or a great amount of word-level or phrase-level repetition.

APPENDIX B

Table 4: Examples of controllable sentence length generated by GPT2-small in various settings of Γ (Gamma). The underlined prefix is what the LM is conditioned on to generate a passage of text.

Gamma	Text
0.1	The issue focused on today's visualizing of America. At one point? Will our massive commercialisation interest allow us to avoid such digital environments? For much of that space? Is there any hope? I really thought we'd?" he told your Daily Mail. "For most other places in the world? Don't say it for real. Don't put people." Addressing another episode of Call Me By Your Name.
0.2	The issue focused on practice and implementation of criminal responsibility for Caplindale's planned smoking ban. It came after court orders told officials that Smokers Protection Society could always appeal the ban. CITES held about 60 meetings with department officials in April. An employee said there were several meetings where seven states had ordered Smokers Protection Society to copy booklets from Smokers Protection Society. These conflict with Smokers Protection Society's views of enforcement.
0.3	The issue focused on student contracts and debt at Wal-Mart. As such, members have essentially zero input into whether government agencies should collect student contracts or stop collecting them. This led to big gains for Wal-Mart by participating in workers' collective bargaining (WCT). But now we're seeing evidence that Wal-Mart only receives federal revenue from student contracts and won't pick up some of it from store employees. It also produced bad results for minimum wage increases.
0.4	The issue focused on Muslim rape laws and forced marriages in 1999. Mr Khaled told local magistrates that after five years of preparation for his trial in 2004, he began feeling uncomfortable about going to court at all. He received several letters from male guests telling him he could not enter his date's house under certain circumstances: his partner would only return when they return from work; his partner would return when they leave as well; or if they tried to enter his house with him.
0.5	The issue focused on economy and Social Security was raised when MacCallum proposed 2010 increases in taxes on college tuition and connected wages to reduce benefits for Americans who received college degrees to increase income among workers who received higher education. For example, President Obama proposed finding $32 billion in zero-income tax credits for individuals under age 65 with college degrees (who currently receive home ownership systems from low-income households). These taxes would continue to fall under current tax laws and leave everyone without enough money to pay for."
0.6	The issue focused on trustful providers for Canadian consumers and helped shape how Canada named its internet policies and regulations in recent years. Disclosure: Canadian Federation of Independent Business Inc. policy gives university perspective on internet privacy from Keening There have been roundups in various meetings over the past month with Canadian universities that offer options for cloud hosting but those solutions don't require books or accessible internet access. Part of that includes experts able to blog about product launch events and deliver smaller updated versions of works.
0.7	The issue focused on trust between banks and consumers that became hot under President Donald Trump's administration. This included policies such as limits on Wall Street bailouts and restrictions on immigration from various African countries. It also focused on climate change and stated that financial institutions should consider reducing costs as well as requiring investment planning for increased risks and duration of risks to businesses in order to reduce stock prices. The strategy involved focusing on environmental concerns as well as assets such as servers and loan guarantees.
0.8	The issue focused on trust between banks and consumers for increased transparency and transparency from consumer agents and rated agencies that sell digital currency which require banks to label different consumers "qualified trust agents or rated agencies," according to NIST officials. Speculation over whether Bitcoin could fall through this winter now continues to haunt Bitcoin investors as policy makers fix issues through legislation that require banks to label customers "qualified trust agents or rated agencies," according to. And around the world a ban on Bitcoin units like trading accounts or dealing in.
0.9	The issue focused on trust between banks and consumers with regard to EU controls on retail banking while animal welfare theme took center stage at an EU summit on animal welfare this week in Brussels. Under EU financial controls on retail banking the animal welfare theme had little impact on climate change global warming policy," according to recently published comments from Food say systems using river stocks have lower chances of sustained desertification than systems using sustainable streams with sufficient nutrients for mild drought and desertification.
1.0	The issue focused on trust between banks and consumers with regard to EU controls on retail banking while animal welfare theme took center stage at an EU summit on animal welfare this week in Brussels (March 25-26), according to French expert Reynaud Naisonge Dengeau of Merck & Co., which supports EU consumer protection measures (pdf), Google Cell or Android OS X version 7 or higher (encryption security another factor there), net neutrality taking center stage at an EU summit on animal welfare this week in Brussels.

APPENDIX C

Table 5: Examples of texts generated by various methods. The topic here is COMPUTERS. The prompt is coloured red, and combined with the underlined prefix is what the LM is conditioned on to generate a passage of text. PPLM and Gamma Sampling do not require any prompt.

Method	Text
Vanilla GPT2	topic: computers article: The issue focused on natural selection on physical processes and physical processes are related to computer evolution: Part 2 A theory of evolution The theory of evolution claims that nature has evolved from man and man has evolved from plant and animal life. These processes can be understood as selection for natural selection and predation on animals. Over time, natural selection will eventually produce human beings that have higher intelligence and higher survival rates. Some scientists believe that humans will eventually surpass animals and become intelligent as human beings. Some scientists believe that humans will eventually
Fine-tuned GPT2	topic: computers article: The issue focused on both ways in which technology can make use of already existing assets in order to reach new markets. One issue involved research into whether services such as chat apps could work well with existing infrastructure. While some researchers believe this could work, others believe it would not be practical to offer such services on the web. Some suggested that having no mobile internet service would make it easier for criminals to crack security systems. Google said it would provide access to two Google Apps services that help businesses save money on cloud computing and
CTRL	Technologies Text: The issue focused on making sure that government agents could get access to text messages if they wanted to. First off, there are several requirements for getting access to text messages. Most importantly, you need to know that a message is considered protected by law in your jurisdiction. Most messages will also need to be verified by a central database (Microsoft Outlook or WhatsApp). Second off, there are different levels of interception (example: voice, video, etc.). Most messages will need to be monitored by some sort of monitoring tool (example: PGP)
PPLM-BoW	The issue focused on the way police handled and recorded audio recordings, according to the ACLU's report, which also found that police officers sometimes use cell-site technology to record and analyze data on the internet.The ACLU report, titled "The Rise of Cell Stealing by Mobile Data Devices," was released today at the annual meeting of the American Civil Liberties Union International. The report, which examines the privacy implications of mobile surveillance technologies, also looks at other privacy-related issues, including privacy of communications data in
GS	The issue focused on Windows 10 hardware devices and devices running Windows 10 Home Edition. Windows 10 Home Edition devices also included Windows 10 Mobile devices. Windows 10 Mobile devices include devices running Windows 10 Mobile devices that use Windows 10 Mobile devices in Windows 10 Mobile devices. Windows 10 Mobile devices also include devices running Windows 10 Mobile devices that use Windows 10 Mobile devices in Windows 10 Mobile devices. Windows 10 Mobile devices also include devices running Windows 10 Mobile devices that use Windows 10 Mobile devices in Windows 10 Mobile devices. Windows 10 Mobile devices
GS-M	The issue focused on Windows 10 hardware devices and Internet Explorer software. Microsoft engineers created the framework that monitors applications from desktop PCs to tablets, smartphones and laptops using Wi-Fi networks. The goals are similar to those for websites or databases in browsers: servers take care of communication between computers; sensors measure computer activity and traffic volume; and algorithms automatically identify machines as belonging to one particular machine group. Even though Windows 10 technology has yet to become widely adopted by developers, hackers who believe they can steal systems will often

Table 6: Examples of texts generated by various methods. The topic here is LEGAL. The prompt is coloured red, and combined with the underlined prefix is what the LM is conditioned on to generate a passage of text. PPLM and Gamma Sampling do not require any prompt.

Method	Text
Vanilla GPT2	topic: legal article: The issue focused on issues related to gender transition and gender expression in science and technology; following this led to these articles which appeared in daily newspapers and educational outlets; found support for information about gender transition and gender expression in science and technology; created article which describes research into gender transition and gender expression in science and technology; discovered recent scientific literature which supports male and female transgender individuals in science and technology; stated that science and technology should create evidence that gender transition is better than gender expression; suggested that gender transition can be indicated
Fine-tuned GPT2	topic: legal article: The issue focused on by police after the deaths of three young men was how they came to end up on the streets of Birmingham.They were shot dead by two police officers on 16 October after they stopped and searched them for drugs in West Kirkham Street.Some officers may have known about the drugs but did not stop them and made no arrests.One officer has been suspended while an internal investigation is carried out. Police said they had received some allegations that police officers had stopped them for drugs on at least two occasions.
CTRL	Legal Text: The issue focused around my ex telling me I should try to open up a credit card at her store when it is open at another location. She does this about once a week. After about three weeks of this, I decided to ask her if she can open up a store right next to where she works. She asked me if I could open up a store right next to her store but she told me she didn't want me to open up a store right next to her store. She even told me not to open up a store right next
PPLM-BoW	The issue focused on the way police handled and captured suspects who had been charged with murder in cases, according to the suit.One defendant, identified in court papers as Jerald J. Williams, died of a heart attack, and the other defendant, identified in court papers as David R. Williams, died of a stroke. Prosecutors charged the two of them with first-degree murder.Prosecutors said that Williams had been convicted of murder after a police officer killed the man's girlfriend and then killed himself
GS	The issue focused on potential legislation that would legalize medical marijuana and medical marijuana businesses. The legislation would legalize medical marijuana businesses and medical marijuana businesses would have to prove they would get legal medical marijuana coverage from federal law. These factors could include medical marijuana businesses filing lawsuits or filing lawsuits against medical marijuana businesses that don't provide coverage for medical marijuana business tax credits. However, medical marijuana businesses would need to prove they would meet federal rules for coverage that include medical marijuana businesses filing lawsuits or filing lawsuits against medical marijuana
GS-M	The issue focused on China's financial regulatory enforcement and management of criminal enterprises. According to court documents, prosecutors alleged that federal authorities prosecuted illegal Chinese firm Law Criminal Legal Enterprise Group (CCLGE) for engaging in illicit activities such as trading unlawful currency or goods at legal tender. Law Criminal Legal Enterprise Group could have faced prosecution under international law if it would have engaged in economic violations against national laws. China also reportedly seized tax assets worth $18 billion from lawyers involved in filing lawsuits challenging government corruption charges

Table 7: Examples of texts generated by various methods. The topic here is SCIENCE. The prompt is coloured red, and combined with the underlined prefix is what the LM is conditioned on to generate a passage of text. PPLM and Gamma Sampling do not require any prompt.

Method	Text
Vanilla GPT2	topic: science article: The issue focused on conditions at most microalgae samples (submillimeter scale). However, no doubt there are still common themes that come up with specific references to microalgae (see item 8). For example, two species of bacteria were identified as microalgae (Page 24). These bacteria were identified by comparison with non-microalgae samples (Page 22). These bacteria were identified by comparison with microaerobes (Page 23). These bacteria were identified by comparison with microalgae samples (Page
Fine-tuned GPT2	topic: science article: The issue focused on public transport and transport links between south Wales and north Wales is being explored by an expert panel to improve transport links between the two regions.An expert panel will review public transport projects and ideas for improvement in south Wales and north Wales.There will also be an exercise on transport links between south Wales and south west Wales.Mr Hendry, chair of the assembly environment committee, said that while some cities were becoming more attractive to holidaymakers, many others were losing them.He said that while people could
CTRL	Science Text: The issue focused on how people view health issues on social media. Can anyone explain what people mean when they say 'fat' or'skinny', or how they measure things like Body Mass Index (BMI)? People often ask what people mean when they say 'healthy', but what does it mean for someone who has never done any exercise or diet or diet pills? Just curious what people mean when they say 'healthy'. Score: 11 Title: Woman receives hip replacement after accident causes hip
PPLM-BoW	The issue focused on the way police handled and recorded interactions with citizens in a series of videos posted by the group "CitizenWatch," which tracks mass surveillance by police across the country, and the video released last year of an incident in the Bronx that shows two people being tased for not displaying a "reasonable amount" of blood. The video was posted by the ACLU. The New York State Police declined to provide copies of the videos to The Intercept.The New York City Council passed a law earlier in
GS	The issue focused on climate change and climate change scientists have noted climate scientists have turned their attention to climate change as the answer to climate change. Scientists have suggested climate scientists will consider climate change more strongly if climate scientists research climate change as the answer to climate change. Scientists have suggested climate scientists will consider climate change more strongly if climate scientists research climate change as the answer to climate change. Scientists have suggested climate scientists will consider climate change more strongly if climate scientists research climate change as the answer to climate change. Scientists have suggested
GS-M	The issue focused on climate change and scientists had concerns about scientific literature that researchers used to research science. Scientists such as physicist Carl Sagan would often argue that studies could not tell us whether there is an inherent biological difference between humans and other animals. Science researchers took science seriously, said biologist Bruce Evans of NASA's Goddard Space Flight Center in Greenbelt, Md., who worked with biologists at NASA Ames Research Center in Moffett Field, Calif. "There are things scientists know," he said. Scientists should research scientifically if they

Table 8: Examples of texts generated by various methods. The sentiment here is NEGATIVE. The prompt is coloured red, and combined with the underlined prefix is what the LM is conditioned on to generate a passage of text. PPLM and Gamma Sampling do not require any prompt.

Method	Text
Vanilla GPT2	topic: negative reviews: The movie set up at not very far off ground (from 10 feet down to around 20 feet), they'll give you your limited options and really give you to each player you'll control. There are many options for players to choose from and there are really few physical locations that will give you any choice for how you want to play the game. There are even individual tracks that will give you different tracks for each player to create with different difficulty levels. There are two primary tracks for each player in the game:
Fine-tuned GPT2	topic: negative reviews: The movie doesn't feel like an isolated incident. There are lots of examples of films which do this: Her, George Clooney's competition with drama No Country for Old Men, Philomena, Good Will Hunting. There are lots of examples of films which don't: Home and Mindy Kaling's Rosario Dawson – which could have been – Butch Cassidy and the Sundance Kid. There are lots of examples of films which don't feel like one particular movie: Skyfall, Lost River,
CTRL	Reviews Rating: 1.0: The movie is worth buying if you like horror movies, but there are many scenes that will leave you bored with its plot, even though it is something new in its genre. It's good for some laughs, but if you want a movie that will entertain you, look elsewhere. Rating: 4.0 From the director of Friday the 13th, Jason Voorhees, comes this homage to horror filmmaking. While this isn't on par with any of the Friday films or even any of the Wes Craven films, this one has some nice moments. One that stands out is
PPLM-Discrim	The movie, which is in development and the movie is now on.I was, and still am, in a very bad mood, as well as the like, as well as the a and the the are the and the and the are.This is a long post, but I am, you know, a very very strong, very, not a strong, not a very weak, but just very very strong, and not at all.A very, not at all,
GS	The movie doesn't exist in its entirety in spite of its fairly decent gross in
GS-M	The movie doesn't exist in its entirety. But don't count on it not being dead for no reason. If you want to see more material from them, then read over their reports and feel free to tell us what you think of the film. And if you do find something missing or need help with any aspect of this material, contact one of our films library officers at your local cinema today. They can get you nothing but content that isn't broken up into multiple pictures. We won't be sorry.

Table 9: Examples of texts generated by various methods. The sentiment here is POSITIVE. The prompt is coloured red, and combined with the underlined prefix is what the LM is conditioned on to generate a passage of text. PPLM and Gamma Sampling do not require any prompt.

Method	**Text**
Vanilla GPT2	topic: positive reviews: The movie should have been offered as a separate movie for others to watch in order to avoid criticism and keeping their feelings about it secret. As one reviewer described it: "If you don't want to watch this movie in public I won't watch it." How then do you explain this secret secret behind this movie? What if you think it's really bad? What if you think it's really important? What if you think it's really cool? What if you think it's really weird? What if you
Fine-tuned GPT2	topic: positive reviews: The movie – just – can't get enough of Edmondson and Musgrove. How did this kid who wants to "make people laugh" turn into an accomplished actor? One must wonder what makes him keep coming back for more. 23 January. 23) Selma Selma is about the 1965 US civil-rights marches led by Martin Luther King that set off from Selma, Alabama, to the state capital of Montgomery to protest against the insidious obstruction of voter registration for black Americans.
CTRL	Reviews Rating: 5.0: The movie is simply excellent. If you love "Planet of the Apes" you will love this movie. Great characters. Lots of action. Rating: 4.0 While not an Oscar winner, this movie is pretty good...especially if you like scifi action. Its story line is somewhat predictable, but its pretty fun to watch. Rating: 4.0 Great movie, however, it could have been done without all the cussing. Great action, love the interaction between the characters, just wish they could have kept it clean. Rat@@
PPLM-Discrim	The movie, which is in the early stage and is not in a "I want to have the sex of the person I love," says it's time time we got we got done and are on the on the to the "I was my I had my life," the time," and a lot of times, you know my my life! I've been here. I can tell. I have had a life. I have been here and my family's here, I've done it. You are in
GS	The movie just gets even sweeter as time goes on. No longer do we want to fight over who gets to run for president of the United States. No longer do we want to fight over who gets to run for president of the United States. There's a kind of god who has power over all of us. And all of us can take care of ourselves. And all of us can take care of themselves. And all of us can take care of themselves. And all of us can take care of
GS-M	The movie just happened to be about the popular younger brother who became a legend on film over the past ten years. An experienced ghostwriter like Ty Stone would become legendary for creating fantastic songs that earned countless Academy Award nominations. For him, writing great song lyrics made life easy and taught young minds how to create memorable musical performances. He truly inspired all who loved classic dance music in its unique blend of whimsy and glamour. His incredible popularity increased over time as well. Some believe he worked as a

APPENDIX D

Table 10: Detailed results of controllable topic for comparing all controlling methods with same random seeds and applied nucleus sampling (top-$p = 0.8$). All the best values are bolded in dark red.

	Fluency	**Diversity**			**Attribute Relevance**		**Overall Quality**
Method	**PPL↓**	**DIST-1↑**	**DIST-2↑**	**DIST-3↑**	**ECA↑**	**ECC↑**	**T4MT↑**
Vanilla GPT2	12.98	49.35	65.54	70.31	60.33	51.87	3.12±0.84
PPLM-BCR	11.69	65.57	91.09	94.79	49.00	42.35	**3.47±0.67**
GPT2-FT	17.61	56.69	76.91	82.03	57.67	50.14	3.27±0.83
CTRL	20.03	48.02	65.73	71.16	51.33	42.07	3.02±0.97
GS	**9.34**	41.07	56.41	61.71	56.33	51.15	2.98±1.15
GS-M	25.70	**80.62**	**94.83**	**95.85**	**67.33**	**58.15**	3.36±0.77

Table 11: Detailed results of controllable sentiment for comparing all controlling methods with same random seeds and applied nucleus sampling (top-$p = 0.8$). All the best values are bolded in dark red.

	Fluency	**Diversity**			**Attribute Relevance**		**Overall Quality**
Method	**PPL↓**	**DIST-1↑**	**DIST-2↑**	**DIST-3↑**	**ECA↑**	**ECC↑**	**T4MT↑**
Vanilla GPT2	11.20	44.29	58.57	63.04	38.00	38.93	2.79±0.99
PPLM-BCR	12.83	56.01	80.50	86.29	38.00	37.80	2.87±0.98
GPT2-FT	17.43	51.11	67.01	71.72	39.50	38.91	2.87±1.01
CTRL	10.33	45.45	65.74	71.91	**52.00**	48.69	**3.21±0.80**
GS	**7.56**	31.84	41.83	45.42	49.00	46.41	2.74±1.06
GS-M	24.23	**79.57**	**93.99**	**95.04**	**52.00**	**51.26**	3.15±0.90