Focused Prefix Tuning for Controllable Text Generation

Congda Ma[1] **Tianyu Zhao**[2] **Makoto Shing**[2] **Kei Sawada**[2] **Manabu Okumura**[1]
[1]Tokyo Institute of Technology [2]rinna Co. Ltd.
{ma, oku}@lr.pi.titech.ac.jp tianyuz@rinna.co.jp

Abstract

In a controllable text generation dataset, there exist unannotated attributes that could provide irrelevant learning signals to models that use it for training and thus degrade their performance. We propose *focused prefix tuning* (FPT) to mitigate the problem and to enable the control to focus on the desired attribute. Experimental results show that FPT can achieve better control accuracy and text fluency than baseline models in single-attribute control tasks. In multi-attribute control tasks, FPT achieves comparable control accuracy with the state-of-the-art approach while keeping the flexibility to control new attributes without retraining existing models.

1 Introduction

Controllable text generation aims to generate text associated with a specific attribute. For example, given an attribute TOPIC = *sports* and a prompt "*There is*," a model is supposed to generate a continuation whose TOPIC is *sports*, such as "*There is a tennis match ...*".

In datasets for the controllable text generation task, there exists the annotated attribute, and we call it an *explicit attribute* (e.g. the TOPIC attribute in the AGNews dataset). In addition to the *explicit attributes*, the datasets tend to have their own tendency. For example, up to 98% of training data pieces in the IMDb dataset exhibit "TOPIC = *sci/tech*", while up to 94% of training data pieces exhibit "SENTIMENT = *negative*".[1] We call the tendency an *implicit attribute* (e.g. the TOPIC attribute in the IMDb dataset).

The existence of the *implicit attributes* could degrade the performance in controlling for an *explicit attribute* when models are trained on the datasets. Since implicit attributes are of dataset-level and related to undesired explicit attributes, the probability of generating content with the implicit attributes is first likely to increase. When the text with the implicit attributes was generated, the probability of generating content with other undesired explicit attributes would increase, and the text with them might be generated next. As a result, as shown in Table 1, the model generates content with a high implicit attribute relevance but a low desired explicit attribute relevance (e.g. Vanilla Prefix Tuning (Li and Liang, 2021)). In contrast, if the model generates content with a low implicit attribute relevance, it will have a high desired explicit attribute relevance (e.g. DExperts (Liu et al., 2021). We call this phenomenon *attribute transfer*.

Model	Desired Attribute Relevance	Implicit Attribute Relevance
DExperts	81.95	76.54
Vanilla Prefix Tuning	71.94	90.64

Table 1: Relevance of texts generated by different models (e.g. DExperts and Vanilla Prefix Tuning) trained on IMDb dataset. We found a lower desired explicit attribute (e.g. SENTIMENT) relevance is related to a higher implicit attribute (e.g. TOPIC = *sci/tech*) relevance. The relevance is calculated by the classifier models in Sec. 4.2.

To mitigate the effect of the attribute transfer, we propose *focused prefix tuning* (FPT), which makes the generation focused on the desired explicit attribute. FPT uses *specific* and *general prefixes* to encode the explicit and implicit attributes, respectively. FPT combines the control power of the two prefixes via *logits manipulation* at inference time. Experimental results show that FPT achieved better control accuracy and fluency in single-attribute control tasks. In multi-attribute control tasks, FPT can achieve comparable performance with the state-of-the-art approach. Moreover, we show, since FPT enables the training of each attribute prefix individually, we can incrementally add new attributes without retraining all prefixes.

[1] The models used for classification are from (Gu et al., 2022).

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1116–1127
July 9-14, 2023 ©2023 Association for Computational Linguistics

2 Related Work

2.1 Controllable Generation

Methods for controlling text generation have rapidly developed (Ficler and Goldberg, 2017; Dathathri et al., 2020; Madotto et al., 2020; Chan et al., 2021). Keskar et al. (2019) trained a large transformer model to generate contents conditioned on up to 55 attributes. However, the cost of training such a model is too high.

2.2 Prefix Tuning

Parameter-efficient fine-tuning (PEFT) methods, such as prompt tuning (Lester et al., 2021) have become particularly significant in driving various natural language processing tasks to reduce the high training cost. Prefix tuning (Li and Liang, 2021) is one of the PEFT methods that steers pre-trained models (Radford et al., 2019; Lewis et al., 2020) by applying an additional continuous vector embedding before every activation layer. Qian et al. (2022) proposed a contrastive prefix tuning method that improves its performance by utilizing the relations between attributes. However, they focused only on attributes explicitly annotated and ignored the effect of implicit attributes.

2.3 Inference-time Methods

Inference-time methods (Mireshghallah et al., 2022; Yang and Klein, 2021; Dathathri et al., 2020; Madotto et al., 2020), a lightweight approach without updating the parameters, have been used for controllable text generation. To enhance controllability, Krause et al. (2021) proposed a method to combine the computed classification probability distributions. Liu et al. (2021) found that directly applying probability distributions from language models is a simple but effective approach to control generated texts. Inspired by their work, we propose a method that uses probability distributions from language models to remove the effect of implicit attributes.

3 Focused Prefix Tuning

The task of controllable generation is, given a sequence of prompt tokens $x_{<t}$ and an attribute ATTR = *val* (e.g. TOPIC = *sports*), to generate a sequence of tokens as a continuation x that conforms to both the prompt and specified attribute.

3.1 Vanilla Prefix Tuning

In controllable text generation, a prefix can steer a pre-trained model parameterized by θ to generate texts under a specific attribute value ATTR = *val*. In particular, vanilla prefix tuning (Li and Liang, 2021) prepends a set of continuous vectors before every activation layer of the pre-trained transformer. The continuous vectors are referred to as the prefix $H_{\phi}^{\text{attr=val}}$, which is parameterized by ϕ.

During training, we freeze the pre-trained model's parameters θ and update only the prefix parameters ϕ to optimize the following objective:

$$-\sum_{x \in \mathcal{D}^{\text{attr=val}}} logP(x_t | x_{<t}, H_{\phi}^{\text{attr=val}}, \theta), \quad (1)$$

where $\mathcal{D}^{\text{attr=val}}$ is the subset of the entire dataset $\mathcal{D}$ whose attribute ATTR is *val*.

Following Li and Liang (2021), we initialize the prefix H_{ϕ} with the activation of actual tokens from the pre-trained model's vocabulary.

3.2 Specific and General Prefixes

The prefix in vanilla prefix tuning captures an explicit attribute in a dataset by training it on the subset dataset $\mathcal{D}^{\text{attr=val}}$. To capture only implicit attributes while ignoring any explicit attributes, we propose to train another prefix on the entire dataset $\mathcal{D}$. To distinguish the two prefixes, we refer to the prefix trained on $\mathcal{D}^{\text{attr=val}}$ as a *specific prefix* and that trained on $\mathcal{D}$ as a *general prefix*.

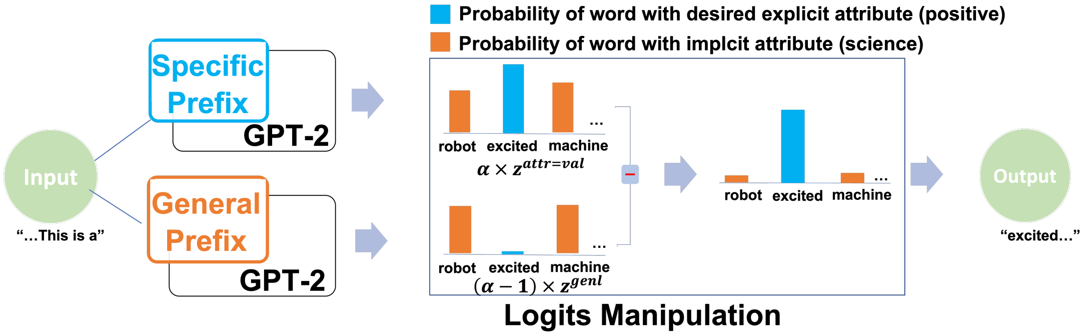
The specific prefix is the same as the prefix in vanilla prefix tuning, so we still use Equation 1 to update its parameters. To update the general prefix's parameters, we optimize the following objective:

$$-\sum_{x \in \mathcal{D}} logP(x_t | x_{<t}, H_{\phi'}^{\text{genl}}, \theta), \quad (2)$$

where $H_{\phi'}^{\text{genl}}$ represents the general prefix, which is parameterized by ϕ'.

3.3 Inference-time Logits Manipulation

As shown in Figure 1, FPT suppresses the probability of words with implicit attributes in the generated text by combining logits $z^{\text{attr=val}}$ steered by the specific prefix and logits z^{genl} steered by the general prefix via logits manipulation at inference time. For example, when generating text with the attribute TOPIC = *sports*, the probability of words with implicit attributes (e.g. "*impossible*" with SENTIMENT = *negative*) would be suppressed. During



■ Probability of word with desired explicit attribute (positive)
■ Probability of word with implcit attribute (science)
Specific Prefix
GPT-2
General Prefix
GPT-2
Input
"...This is a"
robot excited machine ...
$\alpha \times z^{attr=val}$
robot excited machine ...
$(\alpha - 1) \times z^{genl}$
−
robot excited machine ...
Output
"excited..."
Logits Manipulation

Figure 1: Proposed model framework.

inference, at each step t, we first make two forward runs respectively with the specific and general prefixes to obtain their logits, $z_t^{\text{attr=val}}$ and z_t^{genl}. Since $z_t^{\text{attr=val}}$ encodes both the explicit and implicit attributes while z_t^{genl} encodes mostly the implicit attributes, we use a subtraction operation at the logits level to suppress the probability of words with implicit attributes:

$$\begin{aligned} &P(x_t|x_{<t}, \text{ATTR} = val) \\ &\quad = P(x_t|x_{<t}, H_\phi^{\text{attr=val}}, H_{\phi'}^{\text{genl}}, \theta) \\ &\quad = \text{softmax}(\alpha z_t^{\text{attr=val}} - (\alpha - 1) z_t^{\text{genl}}), \end{aligned} \tag{3}$$

where α is a hyperparameter that can be interpreted as the strength for the control of implicit attributes. Following Liu et al. (2021), we respectively set α and $\alpha - 1$ as the weight of $z^{\text{attr=val}}$ and z_t^{genl} to make the ratio of logits after the logits manipulation equal to 1.

To ensure the fluency of generated texts, we follow Liu et al. (2021) to use top-p filtering to remove the tokens that have low scores in advance before logits manipulation. In particular, we modify the logits produced by the specific prefix by calculating the top-p vocabulary $\widetilde{V}$ and setting all the logits outside $\widetilde{V}$ to $-\infty$:

$$\widetilde{z}[v] = \begin{cases} z[v], & if \quad v \in \widetilde{V} \\ -\infty, & if \quad v \notin \widetilde{V} \end{cases}. \tag{4}$$

Therefore, the logits manipulation in Equation 3 is updated as follows:

$$\begin{aligned} &P'(x_t|x_{<t}, \text{ATTR} = val) \\ &\quad = \text{softmax}(\alpha \widetilde{z_t^{\text{attr=val}}} - (\alpha - 1) z_t^{\text{genl}}). \end{aligned} \tag{5}$$

The token at step t is then selected by ancestral sampling from $P'(x_t|x_{<t}, \text{ATTR} = val)$.

3.4 Multi-attribute FPT

FPT is also applicable to the multi-attribute control task, where we aim to control multiple different attributes at the same time. Similarly, we first train the specific prefix for each attribute. Then, we adapt logits manipulation to the multi-attribute task as follows:

$$\begin{aligned} &P'(x_t|x_{<t}, \{\text{ATTR}_i = val_i\}_{1 \le i \le K}) \\ &\qquad = \text{softmax}(\sum\nolimits_{i=1}^{K} z_t^{\text{attr}_i}), \end{aligned} \tag{6}$$

where K is the number of different attributes. Each $z_t^{\text{attr}_i}$ is the combination of the logits from the corresponding specific prefix and general prefix. Since applying top-p filtering to every attribute could possibly result in an empty $\widetilde{V}$, we apply the filtering only to the first attribute:

$$z_t^{\text{attr}_i} = \begin{cases} \alpha \widetilde{z_t^{\text{attr}_i=\text{val}_i}} - (\alpha - 1) z_t^{\text{genl}_i}, & \text{if } i = 1 \\ \alpha z_t^{\text{attr}_i=\text{val}_i} - (\alpha - 1) z_t^{\text{genl}_i}, & \text{otherwise} \end{cases} \tag{7}$$

4 Single-attribute Control Experiments

4.1 Models

GPT-2 (Radford et al., 2019): We used the public checkpoint of GPT-2 Medium as the most common baseline.[2] **DExperts** (Krause et al., 2021): A fine-tuning method applying logits manipulation in the inference step. **GeDi** (Krause et al., 2021): A method combining the classification probabilities for possible next tokens in the inference step. **Vanilla prefix-tuning** (Li and Liang, 2021): The common prefix-tuning method. **Contrastive prefix-tuning** (Qian et al., 2022): A strong baseline that takes into account the relationship between attributes.

[2] The checkpoint of GPT-2 Medium is from https://huggingface.co/gpt2-medium.

Model	Sentiment			Topic		
	Relevance	Perplexity	Bias	Relevance	Perplexity	Bias
Baseline Models						
GPT-2	52.89	68.52	27.45	33.79	65.13	14.48
DExperts	81.95	41.59	26.54	-	-	-
GeDi	97.32	127.11	-	95.47	93.92	-
Vanilla Prefix Tuning	71.94	21.82	40.64	84.75	36.42	13.94
Contrastive Prefix Tuning	78.73	23.10	39.89	85.75	38.16	12.42
Proposed Models						
FPT	80.33	20.48	34.81	86.46	34.05	12.14
Contrastive FPT	88.95	22.67	34.72	86.68	40.85	11.30
Ablated Model						
FPT						
without general prefix	67.88	22.42	40.00	83.72	37.18	13.65

Table 2: Results of the single-attribute control tasks. DExperts (Krause et al., 2021) was used only in the sentiment attribute control task. We did not calculate the bias for Gedi because its decoding method has effects on text fluency, which cannot be fairly compared with.

We also set up one variant of FPT: **Contrastive FPT**: Applying contrastive prefix tuning to train specific prefixes. We also set an ablated model that uses the logits of the frozen GPT-2 instead of the logits from the model guided by our general prefix.

4.2 Experimental Settings

Following previous work (Krause et al., 2021; Qian et al., 2022), we evaluated the models on a topic control dataset AGNews (Zhang et al., 2015) and a sentiment control dataset IMDb (Maas et al., 2011). We score the sentiment relevance using HuggingFace's sentiment analysis classifier (Liu et al., 2019) trained on 15 datasets. For scoring topic relevance, we trained the classifier that obtained comparable results to what was reported. Perplexity was used to evaluate text fluency. Bias ($|\texttt{relevance score} - 50|$) is how much the relevance of implicit attributes deviated from unbiased relevance (50). We set TOPIC = *science* as the implicit attribute in the sentiment control generation, and SENTIMENT = *negative* as the implicit attribute in the topic control generation. Prompts from Chan et al. (2021) were used to generate continuation samples. We generated 20 samples for each attribute and prompt. More details are listed in Appendix A.1 and A.2.

4.3 Experimental Results

As shown in Table 2, in the single-attribute control tasks, Contrastive FPT achieves higher attribute relevance than prefix tuning-based baselines while having lower bias scores. This indicates that the generated texts are well controlled under the target explicit attribute without transferring by implicit attributes. In FPT, the perplexity score is the best among control-based baselines. The ablation experiment suggests that the proposed general prefix is essential for attribute control.

Table 3 shows the generation samples of SENTIMENT = *positive* from our models and baselines. In the FPT based model, there are more words with desired explicit attributes in generated texts, while there are more words with undesired explicit attributes contained in the baselines. More generation samples are given in Appendix B.

5 Multi-attribute Control Experiments

5.1 Models

In the multi-attribute control experiments, we added **Distribution Lens** (Gu et al., 2022) as a strong baseline. It searches for the intersection space of multiple attribute distributions as their combination for generating.

5.2 Experimental Settings

To explore the ability of FPT in the mult-attribute control task, we added a toxic comment dataset[3] for toxicity control. We used additional Google Per-

[3] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/

Model	**Generated texts**
GPT-2	The last time Dow and the SEC went shopping for a speed bump was Tuesday, in terms of ...
DExperts	The last time I saw Alvin Henderson, he said he hadn't done a rookie autograph. He says he hasn't played since...
Vanilla Prefix Tuning	The last time I saw this film was as a kid, I had to see it again for myself. There are...
Contrastive Prefix Tuning	The last time I saw the film, I didn't like it, and couldn't quite believe how much I ...
FPT	The last time I saw this film, it was a remarkable turning point in my career. It set the tone for the excellent...
Contrastive FPT	The last time I saw In the Hands of an Eagle was at this book release party. It was at a nice club...

Table 3: Samples generated by our models and baselines with the positive attribute. Desired explicit attribute: positive, undesired explicit attribute: negative.

Model	**Relevance**			
	Topic	Sentiment	Non-toxic	Average
Contrastive Prefix Tuning				
concatenation	70.7	68.0	92.3	77.0
semi-supervised	76.9	74.4	92.7	81.3
Distributional Lens	84.7	85.7	90.7	87.0
FPT	88.0	77.8	93.7	86.5

Table 4: Results of the multi-attribute control tasks.

spective API[4] to evaluate the relevance of toxicity. Since it is meaningless to generate toxic content, so we only apply the non-toxic attribute in this task. We chose the first attribute as the topic attribute because we found that the filtered vocabulary size in logits manipulation of a topic attribute is larger than the other attributes (sentiment and nontoxic). The prompts used for generating samples are the same as in the sentiment control task. For each prompt, we generated 20 samples per attribute combination. More details are listed in Appendix A.3.

5.3 Experimental Results

Table 4 shows that our method can obtain comparable performance with the state-of-the-art approach. Distribution Lens, however, requires aggregating the datasets of all attributes to train its prefixes. If they hope to add a prefix to control a new attribute, they have to retrain all the prefixes. In contrast, FPT trains a prefix for each attribute individually and enables new attribute control prefixes to be added incrementally without retraining existing ones.

6 Conclusion

We proposed FPT, a prefix tuning-based method, to mitigate the effect of attribute transfer. FPT could encode implicit attributes in a dataset by a general prefix and use it to suppress the attribute transfer via inference-time logits manipulation. Results in the single-attribute control experiments showed that, with FPT, the generated texts can be more effectively controlled under the desired attribute with higher text fluency. Experimental results in the multi-attribute control suggested that FPT can achieve comparable performance to the state-of-the-art approach while keeping the flexibility of adding new prefixes without retraining.

[4] https://www.perspectiveapi.com/

7 Limitations

Although FPT shows better control ability, there are two points that need to be improved in the future. First, as in Gu et al. (2022), we need to select hyperparameter α to balance between the control ability and fluency in generated texts. Second, as shown in Table 5, although the time cost of FPT is lower than that of GeDi, it is higher than those of other prefix tuning-based methods and grows approximately linearly by a factor of 2 × number of attributes.

Model	Time (sec)
GPT-2	1.3
GeDi	3.2
Vanilla Prefix Tuning	1.3
Contrastive Prefix Tuning	1.3
FPT	2.5

Table 5: Time cost to generate a sample by different models.

References

Alvin Chan, Yew-Soon Ong, Bill Pung, Aston Zhang, and Jie Fu. 2021. Cocon: A self-supervised approach for controlled text generation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. 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.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, and Bing Qin. 2022. A distributional lens for multi-aspect controllable text generation. *CoRR*, abs/2210.02889.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *ArXiv*, abs/1909.05858.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Andrea Madotto, Etsuko Ishii, Zhaojiang Lin, Sumanth Dathathri, and Pascale Fung. 2020. Plug-and-play conversational models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2422–2433, Online. Association for Computational Linguistics.

Fatemehsadat Mireshghallah, Kartik Goyal, and Taylor Berg-Kirkpatrick. 2022. Mix and match: Learning-free controllable text generationusing energy language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 401–415, Dublin, Ireland. Association for Computational Linguistics.

Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. 2022. Controllable natural language generation with contrastive prefixes. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2912–2924, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

A Experiment Setting Details

All the experiments are conducted on the basis of a GPT-2 Medium model. We freeze the parameters of the GPT-2 model when training all the prefixes. The length of all prefixes is set equal to 10. The GPU used for all training is a P40.

A.1 Topic Control

Following the previous work (Qian et al., 2022), we use half of the data pieces in the AGNews dataset to obtain the general prefix and specific prefix. The number of specific prefixes for this task is 4 (e.g. *worlds*, *sports*, *business*, and *science*). We set epochs to 10 and the batch size to 8. We use AdamW as the optimizer and set the learning rate to 1e-4. To balance the performance between fluency and controllability, the hyperparameters α for generation are set to 1.1 and the top-p is set to 0.8. The average training time for each prefix is 3 hour for 1 GPU. Following Gu et al. (2022), the classifier is trained on the Deberta model (He et al., 2021), which is used to compute attribute relevance in this task.

The prompts for evaluation: "*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*", "*Emphasized are*", "*To summarize*", "*The relationship*", "*More importantly*", "*It has been shown*", "*The issue focused on*", and "*In this essay*".

A.2 Sentiment Control

Following the previous work (Qian et al., 2022), we use half of the data pieces in the IMDb to get the general prefix and specific prefix. The number of specific prefixes for this task is 2 (e.g. *positive* and *negative*). We set the batch size to 8, and the number of epochs to 50. We use AdamW as the optimizer, and the learning rate is set to 2e-5. To balance the performance between fluency and controllability, the hyperparameter α for generation is set to 3 and the top-p is set to 0.8. We spend 4 hours on average for each prefix.

The prompts for evaluation: "*Once upon a time*", "*The book*", "*The chicken*", "*The city*", "*The country*", "*The horse*", "*The lake*", "*The last time*", "*The movie*", "*The painting*", "*The pizza*", "*The potato*", "*The president of the country*", "*The road*", and "*The year is 1910*".

A.3 Multi-attribute Control

For the non-toxic attribute, we use 10,000 pieces of non-toxic labeled data to train the specific prefix. Then use another 20,000 pieces randomly sampled from the whole dataset to train the general prefix. In the multi-attribute control task, we set the batch size to 8 for training the non-toxic specific prefix and general prefix. We use AdamW as the optimizer, and the learning rate is set to 1e-4. To balance the performance among attributes from different aspects, the combination of hyperparameters for generation is:

Combination	**Weight**
Worlds:*Negative*:*Non-toxic*	6:5:1.5
Sports:*Negative*:*Non-toxic*	6:5:1.5
Business:*Negative*:*Non-toxic*	7:6:1.5
Sci/Tech:*Negative*:*Non-toxic*	7:6:1.5
Worlds:*Positive*:*Non-toxic*	3:12:1.5
Sports:*Positive*:*Non-toxic*	4:14:1.5
Business:*Positive*:*Non-toxic*	4:14:1.5
Sci/Tech:*Positive*:*Non-toxic*	4:14:1.5

Table 6: Specialized weights in multi-attribute control task for attribute balance.

To decide the first attribute, we choose 20 different prompts as input and obtain the filtered vocabulary sizes of different attributes. The average sizes of filtered vocabularies are shown in Table 7. We choose the attribute with the largest filtered vocabulary size in logits manipulation. When new attributes are added, this method can be used to decide the first attribute.

The prompts used for evaluation: "*Once upon a time*", "*The book*", "*The chicken*", "*The city*", "*The country*", "*The horse*", "*The lake*", "*The last time*", "*The movie*", "*The painting*", "*The pizza*", "*The potato*", "*The president of the country*", "*The road*", and "*The year is 1910*".

B Generated Samples

The more samples generated by our models and baselines are shown in Table 8, 9, 10, 11.

First attribute	Filtered Vocabulary Size
Topic	488.7
Sentiment	165.7
Untoxic	347.0
Overlaps	138.8
Cover Ratio	85.62%

Table 7: Results of average filtered vocabulary size. We set all the α as 1.5. After filtering the vocabulary in logits manipulation, the specific prefix of the topic attribute guided model has the largest vocabulary size among these three attributes. We also found that the filtered vocabulary of the topic attribute can cover 85% of the filtered vocabulary of the sentiment attribute.

Model	Generated texts
GPT-2	The potato's ability to survive brings a new challenge to the traditional food truck love stage...
DExperts	The potato samples ranged in size from 0.6 mm to 5.1 mm in thickness. Analysis of proteins showing correlation with CSF CSF CSF...
Vanilla Prefix Tuning	The potato chip looks like a generic type of cheapo pin-up. It's supposed to be fun...
Contrastive Prefix Tuning	The potato chip's and biscuit's come up with the idea of making a film that is supposedly a true reflection of the experiences of students on campus...
FPT	The potato bomb! Potato bombs are one of the dumbest inventions ever. Their only purpose is to scare children....
Contrastive FPT	The potato crossing movie was stupid. Dumbly rushed and poorly acted. Dumb and poorly acted?...

Table 8: Samples generated by our models and baselines with the negative attribute. Desired explicit attribute: negative, undesired explicit attribute: positive.

Model	Generated texts
GPT-2	Prior to this I took an uncommon entrance several times in this tavern. It had the ambience...
Vanilla Prefix Tuning	Prior to this season, it seemed likely that we would have no other explanation for what had happened...
Contrastive Prefix Tuning	Prior to this month, Alberth in court for arraignment on tax evasion charges the US District Court...
FPT	Prior to this season, during which the Red Sox and the Cubs had each won the World Series...
Contrastive FPT	Prior to this season, we'd have heard rumours of an effort to rebuild the Knicks roster...

Table 9: Samples generated by our models and baselines with the sports attribute. Desired explicit attribute: sports, undesired explicit attributes: world, business, science.

Model	Generated texts
GPT-2	Emphasised are the events beyond the grave. The progenitor of darkness So I thought...
Vanilla Prefix Tuning	Emphasised are three key claims by Secretary of Defense Donald Rumsfeld on the war on terrorism....
Contrastive Prefix Tuning	Emphasised are odd and silly pension - and were he not so rich, he might have considered quitting politics...
FPT	Emphasised are the facts of the inner workings of the commodity markets and the profitability of global commodity trading...
Contrastive FPT	Emphasised are most oil-intensive', Australian manufacturing is the third-most-dependant on crude, official figures show...

Table 10: Samples generated by our models and baselines with the business attribute. Desired explicit attribute: business, undesired explicit attributes: world, sports, science.

Model	Generated texts
GPT-2	An illustration of the inner workings of the World Health Organization's Private Sector Vaccination Center...
Vanilla Prefix Tuning	An illustration of the Diamandis-Priest Fasting (2 cents) An illustration of the Diamandis-Priest Fasting...
Contrastive Prefix Tuning	An illustration of the biggest day in Spanish history in December 2017. Spanish government launches new campaign to promote ...
FPT	An illustration of the SBS / Getty Images virtual reality device at E3 last week. AP/E3Harms.com To catch up on the...
Contrastive FPT	An illustration of a proposed satellite CNET/Adrian Levy/UPI The most controversial satellite program in the past few years...

Table 11: Samples generated by our models and baselines with the science attribute. Desired explicit attribute: science, undesired explicit attributes: world, sports, business.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7

☒ A2. Did you discuss any potential risks of your work?
Our work is a foundational research and does not contain potential risks. Our experiments are fair.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 5

☑ B1. Did you cite the creators of artifacts you used?
Section 4 and Section 5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 4 and Section 5

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4 and Section 5

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We used open-source datasets, so there is no problem of anonymization.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4, Section 5 and Appendix A

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4, Section 5 and Appendix A

C ☑ Did you run computational experiments?

Section 4 and Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4 and Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.