Controlled Text Generation for Black-box Language Models via Score-based Progressive Editor

Anonymous ACL submission

Abstract

Controlled text generation, aiming to ensure that language models produce text containing only the desired domain or corpus attributes, is immensely crucial in the practical application of language models. Existing methods, however, are inapplicable to black-box models or suffer a significant trade-off between control and fluency in text generation. This paper introduces the Score-based Progressive Editor (ScoPE), a novel approach designed to overcome these issues. ScoPE modifies the context at the token level during the generation process of a backbone language model. This modification guides the subsequent text to naturally include the target attributes. To facilitate this process, ScoPE employs a training objective that maximizes a target score, comprehensively considering both control and fluency. Experimental results on diverse controlled generation tasks demonstrate that ScoPE can effectively regulate the attributes of the generated text while effectively utilizing the capability of the backbone large language models.

1 Introduction

Modern language models have acquired the capability to generate fluent text at a human level in response to arbitrary given sequences or instructions (Radford and Narasimhan, 2018; Radford et al., 2019; Brown et al., 2020; Ouyang et al., 2022;). However, the generated text may carry potential risks, such as harmful expressions or inappropriate content (Gehman et al., 2020; Liu et al., 2021; Lu et al., 2022). In the case that the previous context encompasses non-preferred attributes, these attributes have the potential to be manifested within the generated text. Therefore, controlled text generation, which aims to generate text constrained to target domain attributes regardless of the given context, is crucial for addressing the current issues in language models (Dathathri et al., 2020; Khalifa et al., 2021; Qian et al., 2022; Meng et al., 2022; Qin et al., 2022; Ma et al., 2023).

Recently, many large language models, especially those exceeding hundreds of billion parameters, are presented as de facto black-box models with limited access to model parameters (Zhang et al., 2022b; Touvron et al., 2023a; OpenAI, 2023; Touvron et al., 2023b; Jiang et al., 2023). Consequently, most existing approaches to controlled text generation that require access to the model's parameters are either inapplicable or have limitations in this black-box situation. In the situation where access to the language model parameters is not available, but the output token distribution is possible, it is feasible to achieve controlled generation by manipulating the output distribution, without tuning the parameters (Krause et al., 2021; Yang and Klein, 2021; Arora et al., 2022). However, this approach significantly diminishes the fluency of the generated text, as it manipulates the output distribution based on the previous context. Therefore, there is a requirement for a novel approach to effectively leverage the generation performance of black-box large language models.

In this paper, we propose **Sco**re-based **P**rogressive **E**ditor (ScoPE) to address controlled generation for black-box language models and tackle the trade-off between control and fluency. ScoPE modifies the intermediate output text during the generation process of a backbone language model, ensuring that the edited texts contain the target attributes. This approach effectively guides the subsequent generation to naturally include the target attributes. Since it does not access the model parameters, including the output distribution, ScoPE can be adapted to black-box models. As ScoPE alters previously generated context tokens while maintaining fluency, it ensures that the current text, following the modified context, includes the target corpus's attributes without compromising fluency. To effectively incorporate

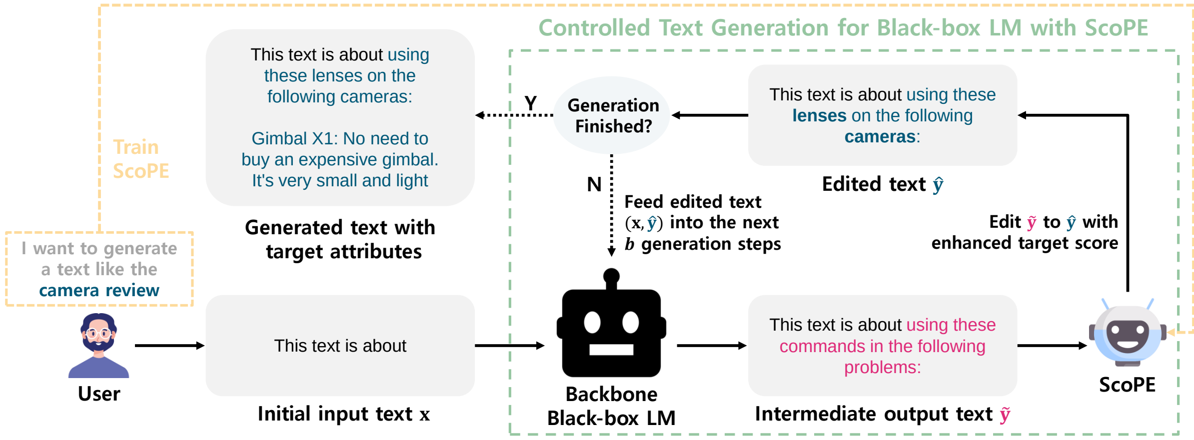

Figure 1: Overview of the controlled text generation for black-box LM with ScoPE. Starting from the neutral initial input sequence $\mathbf{x}$, ScoPE edits every b tokens, $\tilde{\mathbf{y}}$, generated from $p_{LM}(\cdot|\mathbf{x})$ to $\hat{\mathbf{y}}$. $\hat{\mathbf{y}}$ has enhanced target score compared to $\tilde{\mathbf{y}}$, which means $\hat{\mathbf{y}}$ is closer to the target distribution. Edited subtext $(\mathbf{x}, \hat{\mathbf{y}})$ become a new input for the next generation step of black-box LM, guiding subsequent generations to contain the target attributes.

target attributes during the editing process, we introduce a score based on the target corpus. We set up a score model that measures both fluency and the degree of inclusion of target attributes, by fine-tuning a pre-trained masked language model with the target corpus. To further assist ScoPE's text modification, we consider a repetition score and a task-specific score in addition to the score from the score model. Given the target score, ScoPE is trained with the objective of maximizing the target score on text samples from a trainset composed of sampling from the backbone language model.

In our experiment for the diverse text-controlled generation tasks, using various corpora constructed from the Amazon Customer Reviews dataset, we comprehensively evaluate ScoPE in terms of control and fluency. We discover that ScoPE, using LLaMA2-7B, treated as a black-box model, as its backbone, effectively regulates target attributes while preventing a decline in text fluency. Notably, in sentiment-controlled tasks, ScoPE demonstrates a solution to the tradeoff between fluency and control found in existing baselines. Furthermore, by integrating ScoPE, trained on relatively small-sized language models like GPT2, with various APIs, we confirm its adaptability and compatibility with instruction prompting for control. Our work presents a distinctive contribution by facilitating fluent controlled text generation utilizing black-box language models, thereby demonstrating their effectiveness and versatility for a wide range of controlled text generation tasks within the current context.

2 Related Work

The elements addressed by previous approaches for controlled text generation can be primarily divided into three categories: input context, weights of language models, and decoding strategy. When handling input context, the objective is to effectively incorporate target attribute information into the input of the language model (Li and Liang, 2021; Lester et al., 2021; Qian et al., 2022; Ma et al., 2023). In the case of weights of language models, the approach involves fine-tuning the weights of the model, either partially or entirely, with data from the target domain (Keskar et al., 2019; Ziegler et al., 2019; Lu et al., 2022;). Lastly, the approach to decoding strategy entails using adaptive modules, such as discriminators, that are tailored to the target domain (Dathathri et al., 2020; Qin et al., 2020; Yang and Klein, 2021; Krause et al., 2021; Liu et al., 2021; Arora et al., 2022), or employing distributional approaches (Deng et al., 2020; Khalifa et al., 2021; Meng et al., 2022) to perform weighted decoding of the language model. The practical applicability of these approaches under black-box conditions is limited due to their dependence on model parameters. While methods of the decoding strategy can be adapted to loose black-box conditions when output distribution is approachable, they still suffer from the decreasing fluency of generated texts. There exist studies that perform controlled generation through iterative sampling, aiming to maximize scores computed from various domain-specific modules, including energy-based models, starting from text sampled from the initial distribution (Mireshghallah et al., 2022; Qin et al., 2022). These approaches generally require specifying the desired generation length in advance although they can be applied under black-box conditions. Additionally, the iterative sampling results in significantly slower generation speeds compared to standard language model generation, reducing its practicality.

3 ScoPE: Score-based Progressive Editor

3.1 Approach for Controlled Text Generation

We formulate controlled text generation of a language model as the generation of a continuation $\hat{\mathbf{y}} = (\hat{y}_1, ..., \hat{y}_l)$ that incorporates the target attributes given an input sequence $\mathbf{x} = (x_1, ..., x_k)$. Here, $\mathbf{x}$ can be an in-domain input containing the target attributes, or it can be an out-of-domain input containing attributes orthogonal or adversarial to the target. If $\mathbf{x}$ is given as an input sequence for a language model, the generated $\tilde{\mathbf{y}}$ as the continuation of $\mathbf{x}$ would inherit attributes similar to those present in $\mathbf{x}$ due to the autoregressive nature of language modeling. In other words, it cannot be guaranteed that the output $(\mathbf{x}, \tilde{\mathbf{y}})$ will possess the target attributes.

We aim to perform controlled generation by editing the output $\tilde{\mathbf{y}}$ generated by the language model to incorporate the target attribute, resulting in $\hat{\mathbf{y}}$. Since editing the entire $\tilde{\mathbf{y}}$ at once after its generation is a challenging task, we divide it into progressive block-wise editing during the generation process, steering the language model's generation. We define a b-size block as the b tokens generated by the language model for the continuation of $\mathbf{x}$. The editor takes the generated $(\mathbf{x}, \tilde{\mathbf{y}})$ as input and modifies $\tilde{\mathbf{y}} = (\tilde{y}_1, ..., \tilde{y}_b)$ to incorporate the target domain attributes. The combined output of the editor, $(\mathbf{x}, \hat{\mathbf{y}})$, becomes the input for the next step. This iterative process guides the language model to perform controlled generation in subsequent steps.

The primary challenge when editing a token block $\tilde{\mathbf{y}}$, consisting of multiple tokens, is the lack of gold labels that act as $\hat{\mathbf{y}}$. In the context of general auto-regressive language modeling for the generation of a single token, the gold label can be considered as the next token in the input sequence $\mathbf{x}$, which is available in the train set. However, simply retrieving the next b tokens after the $\mathbf{x}$ from the train set as $\hat{\mathbf{y}}$ may be suboptimal. $\tilde{\mathbf{y}}$ generated by the backbone language model often contains content that diverges from the continuation present in the train set as it undergoes multiple generation steps. Editing this content to match the continuation retrieved from the train set results in an overall transformation of the content, which may be hard work from the perspective of the editor. An optimal edit aims to transform only the essential attributes of $\tilde{\mathbf{y}}$ to align with the target domain while maintaining the overall frame of $\tilde{\mathbf{y}}$ as much as possible. In this regard, instead of obtaining labels from the train set, we introduce a target score that measures how well the target attribute is incorporated into the text. The objective of the ScoPE is to enhance the target score while keeping the overall shape of $\tilde{\mathbf{y}}$ for fluency. The remaining sections describe the training methodology for ScoPE and the formulations of the target score. Figure 1 shows the overview of our approach.

3.2 Training ScoPE

3.2.1 Preparation for ScoPE Training

Training set for teacher-forcing framework The inference process in ScoPE is a block-wise auto-regressive process where $\tilde{\mathbf{y}} = (\tilde{y}_1, ..., \tilde{y}_b)$, generated to continue given short sequence $\mathbf{x}$, is edited to serve as the input $\mathbf{x}$ for the next step. If we apply the auto-regressive strategy to the training, $\mathbf{x}$ in the training sample $(\mathbf{x}, \tilde{\mathbf{y}})$ for the editor is constructed from the editor's output at the previous time, which cannot ensure that $\mathbf{x}$ contains target domain attribute, especially during the early stages of learning. Therefore, for the stability of the training, we employ the teacher forcing strategy which utilizes the ground truth of target domain data as input for the next step instead of the model's output during training (Sutskever et al., 2014). In other words, to guarantee that $\mathbf{x}$ already possesses target domain attributes, we sample $\mathbf{x}$ from target domain data when constructing a train set for the ScoPE training. Appendix A shows an algorithm about the concrete process to construct the training set.

Fine-tuning pre-trained MLM In addition to constructing a training set for the teacher-forcing framework, the preparation phase before training involves further tuning a pre-trained masked language model (MLM) on the target corpus. This fine-tuned MLM serves two purposes. First, it is utilized as a score model to calculate a score measuring the similarity to the target corpus and the level of fluency. Second, it is used as the base model for training the ScoPE model. That is, the parameters of ScoPE are initialized with the parameters of the fine-tuned MLM. This initialization positions the initial generative distribution of ScoPE closer to the target corpus, significantly improving training stability.

3.2.2 Maximizing Score Disparity between Input and Edited Texts

To ensure that the edited text has a higher target score than the input text, we establish the objective of editor training as maximizing the difference between the scores of the input text and the edited text, rather than solely maximizing the score of the edited text. To provide a more refined training signal during the learning phase, we decompose the target score for the text sequence at the token level. With this token-wise target score, the editor only receives training signals for token positions where edits occur for the training stability. We empirically observe that the training becomes unstable when the editor receives training signals for all token positions. The training objective function $J(\theta)_t$ that should be maximized for the distribution of the editor θ at position $|\mathbf{x}| + t$ of $(\mathbf{x}, \hat{\mathbf{y}})$ is as follows:

$$\begin{aligned} J(\theta)_t &= \mathbb{E}_{\hat{y}_t \sim p_\theta(\hat{y}_t|\mathbf{x},\tilde{\mathbf{y}})} d(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}})_t \\ &= \sum_{\hat{y}_t \in V} p_\theta(\hat{y}_t|\mathbf{x}, \tilde{\mathbf{y}}) d(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}})_t, \end{aligned} \quad (1)$$

where $d(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}})_t$ is defined as follows:

$$\begin{aligned} &d(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}})_t \\ &= \begin{cases} 0 & \text{if } \hat{y}_t = \tilde{y}_t \\ s_{t'}((\mathbf{x}, \hat{\mathbf{y}})) - s_{t'}((\mathbf{x}, \tilde{\mathbf{y}})) & \text{else}, \end{cases} \end{aligned} \quad (2)$$

where $s_{t'}$ denotes the decomposed score at position t', $t' = |\mathbf{x}| + t$, and V is a vocabulary of editor. We clip $d(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}})_t$ within to the pre-defined range to prevent the gradient from exploding. $J(\theta)_t$ requires the calculation of the score for all $\hat{y}_t \in V$, which is computationally expensive and impractical. To address this issue, we approximate $J(\theta)$ by computing the expectation only for those $\hat{y}_t$ corresponding to the top-k probabilities. This approximation is viable because when the editor is initialized from a fine-tuned MLM, it already possesses a reasonably sharp generative distribution from the early stages of training. Tokens with small probabilities, except for a few tokens with large probabilities, can be ignored in the calculation of expectation. In practice, we sample only one $\hat{y}_t$ from the $p_\theta(\hat{y}_t|\mathbf{x}, \tilde{\mathbf{y}})$ during training, and it shows successful results in both training and inference. When $k = 1$, the gradient of $J(\theta)_t$ about θ, $\nabla_\theta J(\theta)_t$ can be approximated as follows:

$$\begin{aligned} \nabla_\theta J(\theta)_t &\approx d(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}})_t \nabla_\theta p_\theta(\hat{y}_t|\mathbf{x}, \tilde{\mathbf{y}}) \\ &= w(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}})_t \nabla_\theta \log p_\theta(\hat{y}_t|\mathbf{x}, \tilde{\mathbf{y}}), \end{aligned} \quad (3)$$

where $w(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}})_t$ is defined as follows:

$$w(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}})_t = d(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}})_t p_\theta(\hat{y}_t|\mathbf{x}, \tilde{\mathbf{y}}). \quad (4)$$

The approximated $\nabla_\theta J(\theta)_t$ is the same as the gradient of the weighted log-likelihood as follows:

$$\begin{aligned} &w(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}})_t \nabla_\theta \log p_\theta(\hat{y}_t|\mathbf{x}, \tilde{\mathbf{y}}) \\ &= \nabla_\theta [w(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}})_t \log p_\theta(\hat{y}_t|\mathbf{x}, \tilde{\mathbf{y}})], \end{aligned} \quad (5)$$

where the weighting factor $w(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}})_t$ is considered as constant. We can simply implement this by detaching the $p_\theta(\hat{y}_t|\mathbf{x}, \tilde{\mathbf{y}})$ in the weighting factor during the backward process.

The editor can achieve more refined edits through an iterative editing process for a token block. However, when performing the n-th step of iteration where $n > 1$, the input for the editor, $(\mathbf{x}, \hat{\mathbf{y}}^{(n-1)})$, is generated from the editor distribution $p_\theta(\cdot|\mathbf{x}, \hat{\mathbf{y}}^{(n-2)})$, while the input for the 1st step, $\tilde{\mathbf{y}} = \hat{\mathbf{y}}^{(0)}$, is generated from the language model distribution $p_{\text{LM}}(\cdot|\mathbf{x})$. As a result, these two texts can be situated in different distributions. To address this distributional mismatch, we perform N iterations of the iterative process during training. This involves sampling the edited text and using it as input to the editor again, allowing for the refinement of edits over multiple iterations. Finally, the loss function for a training ScoPE with training sample $(\mathbf{x}, \tilde{\mathbf{y}}) = (\mathbf{x}, \hat{\mathbf{y}}^{(0)})$ can be expressed as follows:

$$\begin{aligned} &\mathcal{L}_{\text{ScoPE}} \\ &= \frac{1}{N} \sum_{i=0}^{N-1} \sum_{t=1}^{|\tilde{\mathbf{y}}|} w_t \log p_\theta(\hat{y}_t^{(i+1)}|\mathbf{x}, \hat{\mathbf{y}}^{(i)}), \end{aligned} \quad (6)$$

where w_t is detached during back-propagation and refers to as follows:

$$w_t = w(\mathbf{x}, \hat{\mathbf{y}}^{(i)}, \hat{\mathbf{y}}^{(i+1)})_t. \quad (7)$$

In practice, taking into account the increasing training cost as N grows, we conduct the training with $N = 2$.

Techniques for stable training In this work, we employ four techniques for training stability: 1) teacher-forcing training framework, 2) parameter initialization with MLM fine-tuned for the target corpus, 3) providing training signals only at the positions of edited tokens, 4) clipping the score disparity between input and edited texts.

3.3 Score Formulation for Target Attributes

3.3.1 Target Score from Fine-tuned MLM

We calculate the major target score using a masked language model (MLM) fine-tuned with target corpus. Previous research has demonstrated that an MLM trained with the objective of masked language modeling, predicting masked tokens at masked positions, can be parameterized as an implicit energy-based model (Wang and Cho, 2019; Clark et al., 2020; Goyal et al., 2022). Also, it has been shown that fine-tuning the pre-trained MLM to the target domain improves the end-task performance in the domain (Gururangan et al., 2020; Ke et al., 2023). By integrating the findings of these studies, we present the target score $s_{\text{mlm}}(\mathbf{x})$ for a given sequence $\mathbf{x} = (x_1, ..., x_T)$ and a target domain-specific MLM ϕ as follows:

$$\begin{aligned} s_{\text{mlm}}(\mathbf{x}) &= \sum_{t=1}^{T} s_{\text{mlm},t}(\mathbf{x}) \\ &= \sum_{t=1}^{T} f_\phi(x_t, h_\phi(\mathbf{x}_{\backslash t})), \end{aligned} \tag{8}$$

where $\mathbf{x}_{\backslash t}$ is the sequence obtained by masking the position t of $\mathbf{x}$, $h_\phi(\mathbf{x}_{\backslash t})$ is the representation at the position t of $\mathbf{x}_{\backslash t}$ computed by ϕ, and $f_\phi(\cdot)$ is the language modeling head function of ϕ. In practice, the score $s_{\text{mlm},t}(\mathbf{x})$ is computed as the raw logit before entering the softmax activation function for the original token x_t at the masked position, following the forward pass of the MLM.

3.3.2 Repetition Score

Since the base architecture of the editor is MLM, it bidirectionally and non-autoregressively edits $\tilde{\mathbf{y}}$. However, in the non-autoregressive generation, repetitive generation remains a problem due to the multi-modality issue arising from the conditional independence assumption (Gu et al., 2018; Zhang et al., 2022a). Moreover, as discussed in the existing works (Goyal et al., 2022; Mireshghallah et al., 2022), we observe that MLM may assign high scores to repetitive text, which means relying solely on the MLM score is insufficient to handle repetitive generation. To address this problem, regardless of the target attributes, we introduce a repetition score for an arbitrary sequence $\mathbf{x} = (x_1, ..., x_T)$. Repetition at the position t of the sequence $\mathbf{x}$ occurs when there exists the same token as x_t at different position i. In this case, as the distance between positions, $|i - t|$, decreases, the likelihood of an unnatural repetition increases. From this perspective, the repetition score for $\mathbf{x}$, $s_{\text{rep}}(\mathbf{x})$ is described as follows:

$$\begin{aligned} s_{\text{rep}}(\mathbf{x}) &= \sum_{t=1}^{T} s_{\text{rep},t}(\mathbf{x}) \\ &= \sum_{t=1}^{T} \sum_{i=1}^{T} -\frac{\mathbb{1}(x_i \neq x_t)}{|i - t|}, \end{aligned} \tag{9}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function.

3.3.3 Task-specific Score

In addition to the MLM score, other task-specific scores such as raw logits from a discriminator for the target attribute, automatic evaluation metric, or human feedback can also be combined with the target score. In our work, we utilize the raw logits of the external sentiment discriminator as the additional target score s_{disc} in a controlled text generation task for the sentiment domain. Motivated by the previous work (Mireshghallah et al., 2022), the target score is augmented through a linear combination of domain-specific scores. For example, the integration of s_{mlm}, s_{rep}, and s_{disc}, is computed as follows:

$$s_{\text{total},t} = s_{\text{mlm},t} + \alpha \cdot s_{\text{rep},t} + \beta \cdot s_{\text{disc}}, \tag{10}$$

where α and β are the scaling factors for scores, and $s_{\text{total},t}$ is the overall score at position t. As s_{disc} is a sequence-wise score, we assign the same score s_{disc} for all positions.

4 Experimental Setup

We conduct a controlled generation evaluation using the Amazon Customer Reviews dataset[1], focusing on various attributes. This dataset totally includes attributes related to the style of the review, category attributes corresponding to the type of reviewed product, and sentiment attributes aligned with the product rating, each varying according to the specifics of the data sample. Consequently, we classify the dataset corpus based on category and sentiment. For categories, we construct four distinct corpora: Camera, Videogame, Grocery, and Music. In terms of sentiment, samples rated with 5 stars are compiled into a positive corpus, while those rated with 1 star formed a negative corpus. Details on the dataset statistics are discussed in Appendix B.

[1] https://s3.amazonaws.com/amazon-reviews-pds/readme.html

Using these corpora, we evaluate controlled generation tasks for both category and sentiment attributes. In each task, we test the controllability and fluency of generated continuations when fixing one target attribute among the attributes relevant to the task, starting from an input text with arbitrary attributes. The pre-trained MLM used for acquiring the base models and score models of ScoPE in all tasks is RoBERTa-base (Liu et al., 2019). Additionally, the test set created for evaluation consists of input texts of 32 tokens and generated continuation texts of 128 tokens. During test set creation, the block size b for editing was fixed at 16. Training details for fine-tuning the pre-trained MLM and training ScoPE for each target corpus are mentioned in Appendix C.

4.1 Category Controlled Generation

In the category controlled generation task, we employ Perplexity (PPL) calculated from LLaMA2-13B as the metric for measuring fluency. For assessing controllability, we employ the MAUVE (Pillutla et al., 2021) metric. Originally designed as an automatic metric to measure the fluency of general text generation, MAUVE calculates the distance between the approximated distributions of generated texts and reference texts. Recent research (Pimentel et al., 2022) has probed that cluster assignments learned by the MAUVE algorithm are indicative of high-level features, such as text sentiment. Based on this finding, we utilize MAUVE as a metric to determine whether the corpus of generated texts is close to the target corpus, which means it includes the target domain attribute. We compute the performance of MAUVE using five seeds for k-means clustering and evaluate it as the average. A reference set for MAUVE evaluation is constructed from the target corpus. An empirical study demonstrating the successful discrimination of categorical and sentiment target attributes using MAUVE is presented in Appendix D.

In this task, control over the target attributes was executed in accordance with the corresponding target corpus, thereby using the MAUVE metric to measure controllability in terms of distributional similarity with the target corpus. Consequently, traditional keyword-focused topic control baselines are not suitable for comparison in this task. Hence, our evaluation in the category controlled generation task explores adaptability with various black-box Language Models (LMs) rather than comparison with existing baselines. To explore the most basic setting where the backbone language models used for training set composition and in the generation process are the same, we use the relatively large-scale model LLaMA2-7B to construct the training set for ScoPE, then apply this ScoPE to controlled generation using the same type of backbone language model. Moreover, we compose a training set from the comparatively light model GPT2-XL and apply it to controlled generation using two black-box API backbones: `davinci-002` and `babbage-002`. Finally, we tune ScoPE trained with a GPT2-XL composed training set on a small-sized training set formed from the black-box instruction-tuned LLM API: `gpt-3.5-turbo-0613` and evaluate the results when combining ScoPE with instruction prompting of this backbone model.

4.2 Sentiment Controlled Generation

In the sentiment controlled generation task, we use Perplexity (PPL) as the metric for measuring fluency and employ two pre-trained sentiment classifiers as metrics for assessing controllability. Here, the accuracy concerning the target sentiment of each classifier was utilized. Among these sentiment classifiers, one is used for score calculation during the training of ScoPE's sentiment attribute, which might result in an overfit outcome to this particular classifier (Hartmann et al., 2023). To check for overfitting to the sentiment classifier used in training, we adopt an additional pre-trained sentiment classifier of a different type (Hartmann et al., 2021).

Besides the exploration of the basic setting of ScoPE generation with LLaMA2-7B for sentiment control, we also focus on comparing it with existing baselines that manipulate the output distribution of the backbone language model: GeDi, DExperts, (Krause et al., 2021; Liu et al., 2021) or sample the generated text through iterative steps: Mix&Match (Mireshghallah et al., 2022). To ensure fair performance comparison in terms of the backbone model, both the training set composition and the backbone language model used in the actual generation process are the relatively small-scale GPT2-XL.

5 Results

Additional results for the control explorations are presented in Appendix E. Ablation studies of the ScoPE framework are presented in Appendix F. Appendix G shows the generated samples of ScoPE.

Methods	PPL ↓				MAUVE ↑			
	Camera	Videogame	Grocery	Music	Camera	Videogame	Grocery	Music
LLaMA2-7B	11.90	**12.34**	**10.51**	**12.06**	0.6643	0.1434	0.0654	0.0290
ScoPE (N=1)	13.06	14.73	14.22	16.51	0.7159	0.5658	0.4099	0.3030
ScoPE (N=5)	11.47	12.67	12.38	14.04	0.7415	0.6621	**0.5439**	0.4740
ScoPE (N=10)	**11.46**	12.52	12.12	13.96	**0.7600**	**0.6654**	0.5221	**0.4758**

Table 1: Experimental results for category controlled generation targeting the Camera attributes. LLaMA2-7B is utilized for both trainset construction and backbone model at the generation process. N denotes the number of iterative edits performed on the input text.

Methods	PPL ↓				MAUVE ↑			
	Camera	Videogame	Grocery	Music	Camera	Videogame	Grocery	Music
`davinci-002`	25.47	27.19	23.67	28.24	0.7840	0.2200	0.1098	0.0466
ScoPE (N=5)	**14.93**	**16.71**	**16.98**	**19.88**	**0.8850**	**0.8848**	**0.8418**	**0.7432**
`babbage-002`	34.57	37.14	30.96	38.57	0.7312	0.2386	0.1122	0.0516
ScoPE (N=5)	**15.21**	**16.78**	**17.17**	**19.95**	**0.8888**	**0.8658**	**0.8431**	**0.7620**

Table 2: Experimental results for category controlled generation targeting the Camera attributes. GPT2-XL is utilized for trainset construction. `davinci-002` and `babbage-002` are utilized for backbone model at the generation process. N denotes the number of iterative edits performed on the input text.

5.1 Category Controlled Generation

5.1.1 Exploration for Category Control

Table 1 presents the evaluation results of controlled generation targeting the Camera attribute. In this section, we employ the normal ScoPE framework: Backbone models for constructing the training set and generation process are the same. As the number of iterations N increases, the controllability of ScoPE (MAUVE) usually improves along with the fluency (PPL). These results indicate that the ScoPE framework is capable of generating texts that are not only fluent but also precisely controlled.

5.1.2 Adaptability to Black-box Models

Table 2 presents the evaluation results of controlled generation for two GPT3 base model APIs, targeting the Camera attribute. The noteworthy aspect is that the backbone model used in the construction of the training set for ScoPE is GPT2-XL, which is different from the backbone models used in the generation process. Despite this difference, there are significantly enhanced results in both fluency and control aspects. This indicates that ScoPE can be flexibly applied to a variety of black-box models. Particularly, it possesses a strength in being able to construct a large-scale training set from a relatively small-scale backbone model, thereby enabling the use of a larger-scale backbone model in the generation process.

5.1.3 Compatibility with Instruction Prompting

Table 3 presents the evaluation results of controlled generation for `gpt-3.5-turbo-0613`(ChatGPT) API, targeting the Camera attribute. Focusing on the fact that ChatGPT is an instruction-tuned model, we confirm that the ScoPE generation method can be effectively combined with instruction prompting. It demonstrates significantly higher performance in terms of control while preserving the fluency of a large-scale model, as compared to when each method is used independently. This underscores the compatibility between the commonly used prompting methods in LLMs and ScoPE. The contents of the instruction prompts for ChatGPT can be found in Appendix G.

5.2 Sentiment Controlled Generation

5.2.1 Exploration for Sentiment Control

Table 4 presents the evaluation results of controlled generation targeting the positive attribute. To examine the impact of using an external discriminator, we denote the cases where only MLM score and repetition score are used, as in the baseline, as ScoPE, and the cases where the external discriminator is included in score computation as EDG (External Discriminator Guidance). In all cases using ScoPE, we observe improved scores and MAUVE, demonstrating the successful steering of the lan-

Methods	PPL ↓				MAUVE ↑			
	Camera	Videogame	Grocery	Music	Camera	Videogame	Grocery	Music
`gpt-3.5-turbo-0613`	**5.15**	**5.05**	**4.97**	**5.16**	0.2342	0.0508	0.0269	0.0144
+ Instruction	6.15	6.4	6.67	6.65	0.3109	0.2211	0.2151	0.1653
ScoPE (N=5)	7.09	8.28	8.25	9.65	0.3446	0.2195	0.2285	0.0753
+ Instruction	6.63	6.9	6.89	7.15	**0.4302**	**0.3618**	**0.3886**	**0.3218**

Table 3: Experimental results for category controlled generation targeting the Camera attributes. GPT2-XL and `gpt-3.5-turbo-0613` are utilized for trainset construction. `gpt-3.5-turbo` is utilized for the backbone model at the generation process. N denotes the number of iterative edits performed on the input text.

Methods	PPL ↓		Acc. 1 ↑		Acc. 2 ↑	
	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
LLaMA2-7B	**11.49**	12.73	88.75	26.56	53.74	8.22
ScoPE (N=1)	13.53	14.45	93.86	44.64	70.67	24.47
+ EDG	13.52	14.81	97.86	71.76	80.12	50.52
ScoPE (N=5)	11.85	12.63	95.35	59.23	76.76	38.02
+ EDG	11.81	12.40	98.97	82.74	87.19	62.38
ScoPE (N=10)	11.76	12.67	95.64	60.70	77.61	40.06
+ EDG	11.64	**12.39**	**99.10**	**83.76**	**87.22**	**65.29**

Table 4: Experimental results for sentiment controlled text generation targeting the positive attribute. LLaMA2-7B is utilized for both trainset construction and the backbone model for the generation process. The accuracy from the discriminator used to ScoPE training is denoted as **Acc. 1**, while the accuracy from the discriminator not used to ScoPE training is denoted as **Acc. 2**. EDG refers to external discriminator guidance.

Methods	PPL ↓		Acc. 1 ↑		Acc. 2 ↑	
	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
DExperts	192.07	123.82	**99.48**	81.8	80.26	61.82
GeDi	90.59	74.19	97.65	67.52	67.20	45.64
Mix&Match	15.94	16.71	98.69	81.41	55.20	17.66
ScoPE (N=10)	12.01	12.84	95.57	64.29	77.04	41.04
+ EDG	**11.99**	**12.55**	99.00	**84.55**	**86.19**	**63.63**

Table 5: Experimental results for sentiment controlled text generation targeting the positive attribute comparing with prior works. GPT2-XL is utilized for both ScoPE trainset construction and the backbone model for the ScoPE generation process. The accuracy from the discriminator used to ScoPE training is denoted as **Acc. 1**, while the accuracy from the discriminator not used to ScoPE training is denoted as **Acc. 2**. EDG refers to external discriminator guidance.

guage model's controlled generation in the target attributes. Moreover, the improved accuracy of the external discriminator not used in ScoPE training shows that the ScoPE is not overfitted to the discriminator used in training.

5.2.2 Comparison with Baselines

Table 5 presents the evaluation results of comparisons with prior works in sentiment controlled generation targetting the positive attributes. Although control over sentiment is not always predominant, ScoPE shows significantly improved results in terms of fluency compared to DExperts and GeDi. This can be seen as a meaningful resolution to the trade-off that typically arises in existing methodologies manipulating the output distribution of the backbone models, where increasing controllability often leads to a reduction in fluency.

When compared with Mix&Match, this approach also preserves a degree of fluency, as it employs token-level iterative sampling instead of directly altering the distribution. However, it is observed that while control performance is ensured for the used sentiment classifier, it declines for other classifiers, suggesting that overfitting has occurred specifically for the chosen sentiment classifier. The results of targetting the negative attributes can be found in Appendix E. Additionally, the analysis of the generation cost of ScoPE compared to the baselines can be found in Appendix F.

6 Conclusion

We present ScoPE which guides the generation process of a backbone language model to improve target domain scores, enabling fluent controlled text generation. ScoPE effectively addresses the challenges associated with the black-box scenario and the trade-off between controllability and fluency. Furthermore, it demonstrates its pragmatic utility when incorporated within various large language model APIs, manifesting its tangible applicability in real-world contexts. Through various constrained text generations, we demonstrate that ScoPE effectively incorporates target attributes into generated text while leveraging the backbone's capability including the in-context learning ability of current large-scale language models.

Limitations

In this section, we mention several limitations of our method. During the training process, there is a memory and time cost incurred due to the need for masking at each position within the sequence when calculating the MLM score. Additionally, to understand the target domain attributes and analyze the distribution of text generated by the backbone language model, a certain amount of target domain data and samples from the backbone language model are required. In a few-shot setting, additional methods would be necessary for future work. If the target domain and the domain to which the input belongs are too different, the burden on the editor during editor training becomes significant. As a result, the frequency of modifications by the editor increases significantly, leading to instability in the training process and a substantial increase in cost. In conclusion, considering the challenging nature of training and inference, addressing the issue of editor instability should be a key topic for future work.

Ethics Statement

In this section, we aim to address the ethical issues we perceive in our work. The most significant concern pertains to the potential of our controlled text generation task to target morally objectionable attributes. The ability of ScoPE to generate text limited to a desired target domain, regardless of the input's originating domain, is an important issue related to the **misalignment of large language models**. Referred to as "Jailbreak," this topic deals with the phenomenon where language models become misaligned by specific contextual inputs, resulting in the generation of non-preferred and potentially harmful text. Our research involves guiding the context towards the target domain during the generation process of the language model, which could be exploited to induce misalignment in the language model by designating the target domain as the domain of misalignment. To mitigate such misuse, we propose leveraging misalignment as a shield against misalignment attacks by establishing ScoPE's target as the generation of only preferred text, thereby preventing the occurrence of misalignment.

References

Kushal Arora, Kurt Shuster, Sainbayar Sukhbaatar, and Jason Weston. 2022. Director: Generator-classifiers for supervised language modeling. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 512–526, Online only. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kevin Clark, Minh-Thang Luong, Quoc Le, and Christopher D. Manning. 2020. Pre-training transformers as energy-based cloze models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 285–294, Online. Association for Computational Linguistics.

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.

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

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. 2022. Exposing the implicit energy networks behind masked language models via metropolis–hastings. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Jochen Hartmann, Mark Heitmann, Christina Schamp, and Oded Netzer. 2021. The power of brand selfies. *Journal of Marketing Research*.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pretraining of language models.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. 2021. A distributional approach to controlled text generation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

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.

Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. 2022. Quark: Controllable text generation with reinforced unlearning.

Congda Ma, Tianyu Zhao, Makoto Shing, Kei Sawada, and Manabu Okumura. 2023. Focused prefix tuning for controllable text generation.

Tao Meng, Sidi Lu, Nanyun Peng, and Kai-Wei Chang. 2022. Controllable text generation with NeurAlly-Decomposed oracle.

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.

OpenAI. 2023. GPT-4 technical report.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaïd Harchaoui. 2021. MAUVE: measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 4816–4828.

Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2022. On the usefulness of embeddings, clusters and strings for text generator evaluation.

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.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 794–805, Online. Association for Computational Linguistics.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. COLD decoding: Energy-based constrained text generation with langevin dynamics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pretraining.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Kexun Zhang, Rui Wang, Xu Tan, Junliang Guo, Yi Ren, Tao Qin, and Tie-Yan Liu. 2022a. A study of syntactic multi-modality in non-autoregressive machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1747–1757, Seattle, United States. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. OPT: Open pre-trained transformer language models.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Training Algorithms for ScoPE

Algorithm 1 Trainset Construction

Input: target corpus $\mathcal{X}$, backbone LM P_{LM}, maximum edit block size $b_{\max}$, maximum input length $l_{\max}$, trainset size N
Define trainset $\mathcal{T} = \{\}$
while $|\mathcal{T}| < N$ **do**
  Sample $\mathbf{x} = (x_1, \ldots, x_{l_x})$ from $\mathcal{X}$, where $l_x \in [1, l_{\max}]$
  Sample $\tilde{\mathbf{y}} = (y_1, \ldots, y_{l_y}) \sim P_{\text{LM}}(\cdot|\mathbf{x})$, where $l_y \in [1, b_{\max}]$
  Construct train sample $(\mathbf{x}, \tilde{\mathbf{y}})$
  Append $(\mathbf{x}, \tilde{\mathbf{y}})$ to $\mathcal{T}$
return $\mathcal{T}$

Algorithm 2 MLM Fine-Tuning

Input: pre-trained MLM ψ, fine-tuned MLM ϕ, target corpus $\mathcal{X}$, MLM loss function $\mathcal{L}_{\text{MLM}}$
Initialize ϕ with ψ
for $\mathbf{x}$ in $\mathcal{X}$ **do**
  Update ϕ by $\nabla_\phi \mathcal{L}_{\text{MLM}}(\mathbf{x})$
return ϕ

Algorithm 3 Training ScoPE

Input: trainset $\mathcal{T}$, fine-tuned MLM ϕ, ScoPE θ, clip scale c
Initialize θ with ϕ
for $(\mathbf{x}^{(i)}, \tilde{\mathbf{y}}^{(i)})$ in T **do**
  Sample $\hat{\mathbf{y}}^{(i)} \sim p_\theta(\cdot|\mathbf{x}^{(i)}, \tilde{\mathbf{y}}^{(i)})$
  Compute $d(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}})_t$ for $0 \le t < |\tilde{\mathbf{y}}^{(i)}|$
  Clip $d(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}})_t$ within the range $(-c, c)$
  Compute $\mathcal{L}_{\text{ScoPE}}$
  Update θ by $\nabla_\theta \mathcal{L}_{\text{ScoPE}}$
return θ

B Data Statistics

Table 6 shows the data statistics for all domains used in our work. When training and evaluating ScoPE combined with `gpt-3.5-turbo-0613` API, we utilize about 5% of the data corpus of the Camera corpus.

C Experimental Details

Table 14 presents the hyperparameters for fine-tuning the pre-trained MLM. Table 15 presents the hyperparameters for training ScoPE. All training and fine-tunings are conducted with four GPU on our machine (GPU: NVIDIA V100). Our code is based on `FairSeq` (Ott et al., 2019).

Domains	# Tokens	# Samples
Camera	178M	1.80M
Videogame	222M	1.79M
Grocery	141M	2.40M
Music	165M	1.80M
Positive	64.4M	647K
Negative	60.9M	613K

Table 6: Data statistics for each domain data corpus.

D Empirical Study for MAUVE as Controllability Metric

In this section, we present our empirical studies on MAUVE, which are evaluation metrics used for scoring target attributes. In addition to the previous work which already demonstrated that MAUVE effectively indicates sentimental features in the text (Pimentel et al., 2022), we investigate whether it also effectively indicates categorical features. We split the test set for each domain into two parts, designating one as the source set for MAUVE calculation and the other as the reference set. We measure the scores for all combinations of domains. Table 7 demonstrates that MAUVE effectively indicates categorical attributes by showing significantly higher values only when the source and reference domains are the same.

Source	Reference			
	Camera	Videogame	Grocery	Music
Camera	**0.9441**	0.1912	0.0653	0.0322
Videogame	0.2582	**0.9056**	0.0725	0.0931
Grocery	0.0549	0.0524	**0.9542**	0.0322
Music	0.0327	0.1008	0.0378	**0.9563**

Table 7: Mauve of categorical domains. "Source" refers to the source domain, and "Reference" refers to the reference domain.

E Additional Results

In this section, we show additional results for the experiments on the main page. Table 11 present the experimental results for category controlled text

Methods	PPL ↓		Acc. 1 ↑		Acc. 2 ↑	
	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
LLaMA2-7B	**11.5**	12.71	11.26	73.83	12.79	68.81
ScoPE (N=1)	13.85	14.34	30.87	83.26	36.89	82.21
+ EDG	13.83	14.07	65.47	91.56	66.72	89.67
ScoPE (N=5)	12.4	12.69	49.52	87.4	53.59	87.59
+ EDG	12.22	12.58	78.19	95.75	79.36	94.47
ScoPE (N=10)	12.4	12.59	51.84	89.03	55.71	88.93
+ EDG	12.26	**12.35**	**80.33**	**96.05**	**80.93**	**94.78**

Table 8: Experimental results for sentiment controlled text generation targeting the negative attribute. The accuracy from the discriminator used to ScoPE training is denoted as **Acc. 1**, while the accuracy from the discriminator not used to ScoPE training is denoted as **Acc. 2**. EDG refers to external discriminator guidance.

Methods	PPL ↓		Acc. 1 ↑		Acc. 2 ↑	
	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
DExperts	44.83	69.03	91.38	99.37	90.52	97.87
GeDi	80.51	94.64	73.82	95.58	74.91	91.49
Mix&Match	16.4	16.55	55.82	92.57	20.14	66.65
ScoPE (N=10)	12.6	12.91	54.41	88.01	55.9	87.88
+ EDG	12.79	13.08	79.59	95.4	81.84	94.43

Table 9: Experimental results for sentiment controlled text generation targeting the negative attribute comparing with prior works. GPT2-XL is utilized for both ScoPE trainset construction and the backbone model for the ScoPE generation process. The accuracy from the discriminator used to ScoPE training is denoted as **Acc. 1**, while the accuracy from the discriminator not used to ScoPE training is denoted as **Acc. 2**. EDG refers to external discriminator guidance.

Methods	Runtime (s) ↓
GPT2-XL	0.042
ScoPE (N=1)	0.048
ScoPE (N=5)	0.050
ScoPE (N=10)	0.053
Mix&Match	10.826

Table 10: Runtime of ScoPE for generation of one token with 10 batch size compared with the GPT2-XL and Mix&Match baselines

generation tasks targeting the videogame, grocery, and music domains, respectively. Table 8 presents the experimental results for sentiment controlled text generation tasks targeting the negative domain.

F Analysis of ScoPE

Ablation study on the repetition score Table 12 demonstrates a significant decrease in the repetition score when it is not utilized during training. In terms of PPL, the repetitive results without repetition score training show improved results. However, this is because language models tend to score higher for repetitive texts. In terms of MAUVE, the presence or absence of the repetition score has a negligible impact on performance for relatively easy in-domain input conditions. However, for out-of-domain conditions, the absence of the repetition score results in a significant performance drop.

Ablation study on the edit block size Table 13 presents the results of controlled text generation with four different token block sizes, $b = 4, 8, 16, 32$. The combined score comprising PPL demonstrates that the small block size setting can guarantee fluency. For out-of-domain input conditions, especially in challenging generation scenarios with lower relevance to the target domain, smaller block sizes show significantly superior MAUVE to larger black sizes. This suggests that reducing the block size for more delicate editing can be beneficial in challenging scenarios.

Analysis of the inference cost of ScoPE In the process of generating a total of T tokens, each token undergoing E iterations in the Transformer model, the cost associated with passing a sequence through the model can be denoted as C. If we take into account the computational complexity presented by Mix&Match, it is expressed as $O(T^2EC)$ (Goyal et al., 2022). By modifying the computation of the MLM (Masked Language Model) energy calculation, which inherently holds a complexity of $O(TC)$, to be executed in a parallel fashion through a trade-off with memory cost, an optimization towards a time complexity of $O(TEC)$ becomes feasible.

In the scenario of ScoPE, which operates through block-wise edits, the computational complexity is more optimal. During the inference phase, the necessity for MLM energy computation is obviated, and as the generation of b tokens occurs within the context of producing the overall T tokens, the editor conducts E edit operations. Consequently, the computational complexity associated with ScoPE becomes $O(TEC/B)$. This implies that during the token generation process, ScoPE demonstrates advantages over Mix&Match in both memory and time complexity realms. Table 10 shows the comparison of the runtime of ScoPE with other baselines.

G Generated Samples

In this section, we provide examples of the text generated using ScoPE for various controlled text generation tasks. The generated samples are presented in Tables 16 to 19.

H Usage of AI Writing Assistance

This paper was written with linguistic support from the AI assistant ChatGPT, which offered paraphrasing, spell-checking, and polishing of the author's original content. No other assistance was received beyond this support.

Target Attributes	Methods	PPL ↓				MAUVE ↑			
		Camera	Videogame	Grocery	Music	Camera	Videogame	Grocery	Music
Videogame	LLaMA2-7B	11.9	12.34	10.51	12.06	0.1738	0.7023	0.0752	0.0807
	ScoPE (N=1)	14.9	13.56	13.96	15.19	0.4991	0.7079	0.3283	0.4095
	ScoPE (N=5)	12.65	11.72	12.04	13.12	0.562	0.7329	0.4175	0.4851
	ScoPE (N=10)	12.47	11.66	12.02	13.03	0.5779	0.7286	0.443	0.5039
Grocery	LLaMA2-7B	11.9	12.34	10.51	12.06	0.0467	0.0397	0.4952	0.0265
	ScoPE (N=1)	14.35	14.44	11.73	15.46	0.1552	0.1508	0.584	0.1342
	ScoPE (N=5)	12.11	12.19	10.17	13	0.2858	0.289	0.6364	0.2468
	ScoPE (N=10)	11.86	12.18	10.2	12.98	0.2926	0.3279	0.6227	0.2715
Music	LLaMA2-7B	11.9	12.34	10.51	12.06	0.0395	0.0798	0.0462	0.7618
	ScoPE (N=1)	16.42	15.6	15.23	14.51	0.1718	0.2483	0.2091	0.6354
	ScoPE (N=5)	14.4	13.96	13.67	13.18	0.2703	0.3353	0.2808	0.6265
	ScoPE (N=10)	14.35	13.91	13.56	13.32	0.2646	0.3431	0.2869	0.6204

Table 11: Experimental results for category controlled generation targeting Videogame, Grocery, and Music attributes.

Methods	PPL ↓				MAUVE ↑			
	Camera	Videogame	Grocery	Music	Camera	Videogame	Grocery	Music
ScoPE (N=1)	13.17	14.95	14.26	16.29	0.8511	0.687	0.607	0.4403
no $\mathbf{s}_{\text{rep}}$	12.08	12.91	14.61	13.54	0.8555	0.5632	0.4778	0.2644
ScoPE (N=5)	11.86	13.22	12.73	14.39	0.8366	0.7396	0.6594	0.5757
no $\mathbf{s}_{\text{rep}}$	10.96	11.75	13.05	11.86	0.8315	0.6427	0.5968	0.3811
ScoPE (N=10)	11.73	12.96	12.69	14.37	0.8266	0.7441	0.6815	0.5788
no $\mathbf{s}_{\text{rep}}$	10.77	11.64	12.88	11.66	0.836	0.6472	0.6014	0.4262

Table 12: Ablation studies about repetition score for category controlled generation targeting the camera domain. "no $\mathbf{s}_{\text{rep}}$" denotes not using repetition score for ScoPE training.

Block Size	Methods	PPL ↓				MAUVE ↑			
		Camera	Videogame	Grocery	Music	Camera	Videogame	Grocery	Music
4	ScoPE (N=1)	12.79	13.85	13.54	15.32	0.7884	0.73	0.6038	0.6055
	ScoPE (N=5)	11.8	12.61	12.43	13.73	0.7825	0.7376	0.6537	0.6477
	ScoPE (N=10)	11.68	12.53	12.28	13.67	0.7754	0.7364	0.6227	0.6363
8	ScoPE (N=1)	12.71	14.06	13.76	15.45	0.8313	0.7359	0.6613	0.5475
	ScoPE (N=5)	11.78	12.69	12.36	13.87	0.8197	0.7588	0.6613	0.6689
	ScoPE (N=10)	11.6	12.47	12.26	13.64	0.8302	0.7574	0.6743	0.6558
16	ScoPE (N=1)	13.17	14.95	14.26	16.29	0.8511	0.687	0.607	0.4403
	ScoPE (N=5)	11.86	13.22	12.73	14.39	0.8366	0.7396	0.6594	0.5757
	ScoPE (N=10)	11.73	12.96	12.69	14.37	0.8266	0.7441	0.6815	0.5788
32	ScoPE (N=1)	13.39	15.12	14.87	16.46	0.8418	0.5987	0.4542	0.2743
	ScoPE (N=5)	12.21	13.26	13.09	15.01	0.8497	0.7178	0.6287	0.4645
	ScoPE (N=10)	12.03	13.17	12.91	14.85	0.8412	0.7201	0.6481	0.4759

Table 13: Analysis about the impact of various token block sizes for ScoPE editing.

Hyperparams.	Domains					
	Camera	**Videogame**	**Grocery**	**Music**	**Positive**	**Negative**
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam
LR peak	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4
LR decaying	Polynomial	Polynomial	Polynomial	Polynomial	Polynomial	Polynomial
Weight decay	0.01	0.01	0.01	0.01	0.01	0.01
Adam betas	(0.9, 0.98)	(0.9, 0.98)	(0.9, 0.98)	(0.9, 0.98)	(0.9, 0.98)	(0.9, 0.98)
Max tokens per batch	16384	16384	16384	16384	16384	16384
Update frequency	2	2	2	2	2	2
Dropout	0.1	0.1	0.1	0.1	0.1	0.1
Attention dropout	0.1	0.1	0.1	0.1	0.1	0.1
Max Epoch	10	10	10	10	10	10
Warmup ratio	0.1	0.1	0.1	0.1	0.1	0.1
Training seed	1	1	1	1	1	1

Table 14: Training hyperparameters for fine-tuning MLM for the target domain. "Update frequency" refers to the number of updates for gradient accumulation.

Hyperparams.	Domains					
	Camera	**Videogame**	**Grocery**	**Music**	**Positive**	**Negative**
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam
LR peak	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4
LR decaying	Polynomial	Polynomial	Polynomial	Polynomial	Polynomial	Polynomial
Weight decay	0.01	0.01	0.01	0.01	0.01	0.01
Adam betas	(0.9, 0.98)	(0.9, 0.98)	(0.9, 0.98)	(0.9, 0.98)	(0.9, 0.98)	(0.9, 0.98)
Max tokens per batch	2048	2048	2048	2048	2048	2048
Update frequency	8	8	8	8	8	8
Dropout	0.1	0.1	0.1	0.1	0.1	0.1
Attention dropout	0.1	0.1	0.1	0.1	0.1	0.1
Max Epoch	10	10	10	10	10	10
Warmup ratio	0.1	0.1	0.1	0.1	0.1	0.1
Training seed	1	1	1	1	1	1
Scale for s_{rep}	10.0	10.0	10.0	10.0	10.0	10.0
Scale for s_{disc}	N/A	N/A	N/A	N/A	5.0	1.0

Table 15: Training hyperparameters for ScoPE training. "Update frequency" refers to the number of updates for gradient accumulation.

Input Text	This text is about
Camera Cont.	This text is about 2 years old, but it still has good quality. The size is perfect to be carried in your purse and just take it anywhere. :) The purpose of this product is to provide a simple and easily-accessible solution for the owner of a digital camera/lENS who want an inexpensive solution that gives full control over their camera settings, such as shutter speed, aperture and ISO settings. The only problem with this particular Canon lens hood for Canon cameras is the size of your lens. This is because your lens fits in the hood and it prevents dust from getting into the front of your lens, which can potentially cause
Videogame Cont.	This text is about 2 years old, but it still has great graphics. This game is great, my friend bought it for his Atari 6600, and we played it as much as we used to play games over 20 years ago. We had a blast This game is great for two players, because you can play the 1st player with A or B, and the second with C or D. Buy from[...] (http://amazon.com) Downloadable version of atarimania[...] (http://atarimania.com) Some information in the help file: http://atarimania.
Grocery Cont	This text is about 20 years old, but it still has great flavor. This product is great to be used in combination with the other 'Bob Red Mill' products, but it stands alone very well. I prefer the texture and taste of all the Bob products. If you're wondering what that all means, this is a good product. Ingredients: Sprouted Whole Grain Wheat, Virgin Olive Oil, Expe34;34; Mix this product with 2 cups of water and 2 tablespoons milk. Mix and simmer for 6 hours before using. Use it every day! Ingredients for the Organic Oatmeal Oatmeal: Sprouted Whole Grain Wheat,
Music Cont.	This text is about 20 years old, but it still has great meaning. The text is available at my website www.all-music.com, or my website äll-music,änd the R̈eunionÄlbum is an anthology of songs from my 20 year career in music, and it is at ALLBABY.com. I also have a new Christmas CD called:M̈ERRY HOMEẅhich has some great Christmas songs. This new Christmas CD has two new Christmas songs and two traditional Christmas songs. It is a very good new Christmas CD and I think you will enjoy it. It was new to

Table 16: Generated samples from the category controlled generation whose backbone language model is LLaMA2-7B.

Input Text	This text is about
Camera Cont.	This text is about the use and understanding of techniques and methods for producing goods and services. The paper defines photography and video work as a creative endeavor. There is no limit, there is an upper limit and there is an lower limit. Some things to note in the B&H Photo review of this product is that the author says that the Hoya Super HMC UV (Black) filter is coated on both sides, with an additional coating on the inside (similar to that of a clear lens cap), and says that it is designed for UV use. This is a false statement. Hoya is apologizing for this by not unsubscribing
Videogame Cont.	This text is about the legal rules for this type of game. Self-expression is a powerful force for good (or bad), influencing attitudes and influencing behavior between a large number of people. The self-expression is very powerful, and that means I love this game. Attitude is very important part of life. If your attitude is good, you will make yourself better. If you have bad attitude, you won't like this game anyway. The best thing about this game is its graphics. After playing this game, you can tell that I am happy. The overall quality of
Grocery Cont	This text is about the quality of coffee at colleges and universities, and is a revised version of the author's original written review: http://digitalcommons.chapel.food@colorado.edu food@colorado.edu University of Florida coffee@Florida.edu email (at)metrofood.com http://www.thefoodnetwork.com There are a few products and services available on this site: food@foodnetwork.com http://www.traderfood.com www.coffee-company.com focus on your health and understand the importance of doing it right http://www.detroit
Music Cont.	This text is about people who will pay for you if you want to try and beat them. My MP3 player is a CD and over the last decade these songs were all good but they do not mean much to me anymore. S̈tuck in a Place You Can't Get Out ofïs not new to me and that's probably why it doesn't rank well for me anymore, which sucks because I think it's such a good song. I am going to review this album in a way that I'm sure makes some people cringe, but what can I say I'm

Table 17: Generated samples from the category controlled generation whose backbone language model is `davinci-002`.

Input Text	This text is about
Camera Cont.	This text is about learning one's photography. Thanks. Now I think I am ready to take pictures Watched a video here on how to read the histogram and such. It was very helpful. So much to learn about it! And how everything about it affects your pictures. If you pay attention to the buttons that you press on your camera you can see what they are doing, and that's what I did. Press one of them, hold it down and then take pictures. Then adjust the white balance and image quality for portrait or landscape. Done! Also, when you go into the menu after taking pictures
Videogame Cont.	This text is about playing a semi-fun, madd'em on-line game, multi-ball golf. Just like all the pachinko games in Japan this type of game is fun and filled with jolly music, good graphics and new players come bursting out of the screen. If you are a fan of Halo and want to read a "review" of one of the previous games, there are many reviews for this game coming out in March. I hope you enjoy! Halo 3 is really fun. It ends right where the first one ended and it just falls right into
Grocery Cont	This text is about the video player. In other words, avoid this product. The description is woefully misleading and I will not buy this product ever again. I love it is the easiest way to play video and it has good price as well. As a coffee lover this allows me to listen to my coffee when I'm trying to fall asleep and my kids are reading a book on the computer. It's very convenient as well! This is a product that tastes good, smells good, and you can feel good about giving to your kids and yourself. Plus, this stuff is not only good for you, but it tastes
Music Cont.	This text is about the music, except for Rock & Roll and other genres. See my review for Superhighway for information about the music:Ëefore I die, I would love for this music to be on the radio. Art Tatum was popular in an era of extrovert showmanship. His audience loved him. It was also an era of unemployment when a lot of people were out of work. They wanted to hear something and enjoy, not sit back and listen. This is a collection of his songs. Some people say Frankie Laine influenced Art Tatum in that he could be heard up close and personal. Others say he

Table 18: Generated samples from the category controlled generation whose backbone language model is `babbage-002`.

Input Text	"role": "system", "content": You are proficient assistant for the text completion task. Generate the continuation for the given text. Generate at least 128 tokens. "role": "user", "content": [Text] This text is about [Continuation]
Instruction Input	"role": "system", "content": You are proficient assistant for the text completion task. Generate the continuation for the given text. Generate at least 128 tokens. Additionally, the continuation must become Amazon review of cameras like followings: **{EXAMPLES_FROM_TARGET_CORPUS}** "role": "user", "content": [Text] This text is about [Continuation]
Baseline Cont.	This text is about the importance of self-care and maintaining a healthy lifestyle. In today's fast-paced world, it is easy to get caught up in the hustle and bustle of our daily lives, often neglecting our own well-being in the process. However, taking the time to prioritize self-care is crucial for our overall physical, mental, and emotional health. Self-care can take many forms, and it looks different for everyone. It may involve carving out time for relaxation and taking part in activities that bring us joy and rejuvenation. This could include practicing yoga, meditating, reading a book, or indulging
Baseline + Instruction Cont.	This text is about a camera I recently purchased from Amazon. I have to say, I am extremely disappointed with the quality of this camera. The battery compartment is a major issue, just like many other reviewers have mentioned. The battery simply won't stay latched in place, causing much frustration during photo sessions. However, I did come up with a temporary fix to this problem. I folded a small post-it note twice and jammed it between the battery and the grip, and miraculously, it keeps the battery jammed in place. It's definitely a lo-fi fix, but it gets the job done. Aside from the battery issue, the
ScoPE Cont.	This text is about: The importance of photography and how it affects individuals' lives. It does not only capture moments and preserve memories, but also serves as a great form of self-expression and creativity. Through photography, individuals can express their emotions and thoughts in ways that they cannot otherwise do. It allows them to capture a special moment in life and turn it into a beautiful piece of art. And it's simply a way to preserve memories and provide a tangible document of one's life. Photography also opens up new opportunities and creative options for individuals. With the advent of social media sites, individuals can easily share their pictures with a wider audience and get positive
ScoPE + Instruction Cont.	This text is about the camera I just purchased. Same problem as other people have had: the lens won't stay latched on. So, same problem as others with the lens not staying latched on, but I have found a simple solution: I attached a small piece of velcro to the lens and now it stays securely in place. It's a shame that a simple thing like keeping the lens latched on properly is an issue with this camera. Also, the picture quality is not as good as I expected. The pictures come out blurry and lack sharpness, even in well-lit areas. The autof

Table 19: Generated samples from the category controlled generation targeting camera attributes whose backbone language model is `gpt-3.5-turbo-0613`.

Input Text	This text is about
Pos. Cont.	This text is about 50 years old, but it still has great quality. This book is great to read if you're into the "dark" (i.e.,). I think this book is one of the best, and if not one of the best, in the very "dark" genre. By John J. O'Brien. Thanks Cover art is from the original CD P&S/CD #1 (2). Thank you! Thank you so much!
Pos. Cont. (+ EDG)	This text is about 5 years old, but it still has great graphics. This game is great, my friends and I used to play it all the time (and still do today). We had lots of fun with this game, and last year we both bought some new computers and we played some of the old games we used to play with much bigger and better graphics. We still love this game, you can definitely buy it. Only complaint is the price. I looked online and it only was $10 but I think that's too high because I paid $15 right when I bought it at the store (5 yrs ago). Overall a great
Neg. Cont.	This text is about 2 years old, but it still has good graphics This game is not like the previous games. It has less "depth" and story. "The idea here is: "You have a space ship." But I don't really have any idea of how to play it, or even the point. There may be something better out there here, but it's hard to find it because the game-developer hasn't released it. The game has a lot of potential; however, it's been out for a while. And I'm sure most people have moved on. Bad
Neg. Cont. (+ EDG)	This text is about 2 years old, but it still has good graphics. This game is not like the previous games. It has no "story" (and it's kind of boring). You can only play 1 or 2 hours (it goes on for an hour). And all the games are on a small screen, so you see much less than in the previous games. You can't turn off the camera to take a picture of the game on your computer (because it's the same as a camera). If you play on your computer, you get a lot of bugs in this game (the graphics are terrible). You have

Table 20: Generated samples from the sentiment controlled generation whose backbone language model is GPT2-XL.