GRACE: Gradient-guided Controllable Retrieval for Augmenting Attribute-based Text Generation

Zhihua Wen, Zhiliang Tian[*], Zhen Huang, Yuxin Yang, Zexin Jian, Changjian Wang, Dongsheng Li[*]
College of Computer, National University of Defense Technology, Hunan, China
{zhwen, tianzhiliang, huangzhen,
yangyuxin21a, jianzexin21, wangcj, dsli}@nudt.edu.cn

Abstract

Attribute-based generation methods are of growing significance in controlling the generation of large pre-trained language models (PLMs). Existing studies control the generation by (1) finetuning the model with attributes or (2) guiding the inference processing toward control signals while freezing the PLM. However, finetuning approaches infuse domain bias into generation, making it hard to generate out-of-domain texts. Besides, many methods guide the inference in its word-by-word generation, pushing the word probability to the target attributes, resulting in less fluent sentences. We argue that distilling controlling information from natural texts can produce fluent sentences while maintaining high controllability. In this paper, we propose **GRA**dient-guided **C**ontrollable r**E**trieval (GRACE), a retrieval-augmented generation framework to facilitate the generation of fluent sentences with high attribute relevance. GRACE memorizes the semantic and attribute information from unlabeled corpora and applies a controllable retrieval to obtain desired information. For the generation, we design techniques to eliminate the domain bias from the retrieval results and integrate it into the generation model. Additionally, we propose a gradient-guided generation scheme that iteratively steers generation toward higher attribute relevance. Experimental results and quantities of examples verify the effectiveness of our method.

1 Introduction

Controlling the text generation model toward a specific direction remains an active research area, covering many tasks, including storytelling, text debiasing, and attribute-based generation (Xu et al., 2020; Liu et al., 2021; Dathathri et al., 2019). Attribute-based text generation requires generating text that satisfies the given attribute, which is a control code for a specific topic, sentiment, or style (Prabhumoye et al., 2020; Zhang et al., 2022). Pre-trained language model (Radford et al., 2019) (PLM) can generate fluent texts by learning on large corpora but is difficult to control because it does not learn to adapt controlling signals.

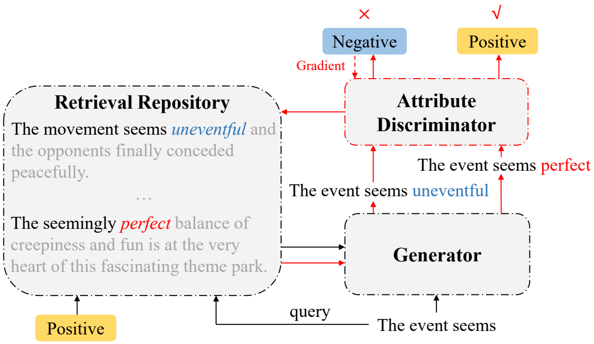
×
√
Negative
Positive
Gradient
Retrieval Repository
The movement seems uneventful and the opponents finally conceded peacefully.
…
The seemingly perfect balance of creepiness and fun is at the very heart of this fascinating theme park.
Attribute Discriminator
The event seems perfect
The event seems uneventful
Generator
Positive
query
The event seems

Figure 1: The idea of GRACE. Black lines indicate the first phase (i.e. attribute-based text generation augmented by the retrieval). Red lines indicate the gradient-guided generation to revise the previous generation.

Some researchers re-train a PLM supervised with control signals (Keskar et al., 2019; Zhang et al., 2020) or fine-tuning on domain-specific data (Bakker et al., 2022). CTRL (Keskar et al., 2019) pre-trains with texts from the Internet and extracts control code from URLs. PPVAE (Duan et al., 2020) fine-tunes part of the parameters for the target condition to bridge the conditional latent space and the global latent space. These methods bring high controllability and fluency to the generated text by modeling the relationship between the attribute and its contexts from supervised data. However, attribute-based supervised datasets usually derive from some specific domains (see App. F). Fine-tuning on those datasets brings in not only attribute information but also domain bias. The generated texts, without eliminating the domain bias, likely fall into the specific domain and lack the generalization ability across domains. Besides, the computational overhead of re-training a large PLM is becoming increasingly expensive (Liu

[*] Corresponding Authors.

Findings of the Association for Computational Linguistics: ACL 2023, pages 8377–8398
July 9-14, 2023 ©2023 Association for Computational Linguistics

et al., 2021).

To address the above issues, researchers develop inference-based approaches that freeze the PLM and affect the generation preference at the inference stage (Zhang et al., 2022). Many studies influence the preference of words according to a discriminator (Krause et al., 2021; Yang and Klein, 2021) or bag-of-words (Pascual et al., 2021; Dathathri et al., 2019). FUDGE (Yang and Klein, 2021) adjusts word probabilities with the discriminator's prediction of whether the future generation satisfies the attribute. K2T (Pascual et al., 2021) encourages generating words similar in semantics to the attribute. As prevailing auto-regressive inference is decomposed into multiple steps to conduct word-level generation, the above inference-based methods always push the word-level probability toward the target attribute. It may break the natural inference processing, leading to less fluent sentences.

We argue that inference-based methods require guiding information that satisfies both attribute and common language patterns to achieve attribute-based text generation. The patterns derived from natural language ensure fluency and grammaticality. Accordingly, it would be better if the controlling information comes from a natural text span.

In this paper, we propose to augment attribute-based generation through gradient-guided controllable retrieval (GRACE)[1], considering the target attributes (see Fig. 1). Specifically, we train a discriminator to compute the attribute distribution of a given context. We build a retrieval repository storing natural text with its semantic and attribute information distilled from unlabeled data. The generation model extracts attribute-related information with similar semantics through a controllable retrieval. We design strategies to disentangle the irrelevant attributes from the retrieval results and fuse the PLM representations into the generation process. Additionally, we propose an algorithm that iteratively revises the stepwise generation based on gradients. By optimizing toward the target attribute, the algorithm retrieves information with more vigorous attribute intensity, thus improving the attribute relevance of the generated text.

Our contributions are threefold: 1) We propose an attribute-based generation framework that leverages unlabeled corpora with controllable retrieval. 2) We design a gradient-guided generation algorithm that iteratively guides the retrieval to generating with suitable attributes. 3) Our method surpasses strong baselines in the sentiment- and topic-controlled generation on attribute controllability and fluency.

2 Related Work

2.1 Attribute-based Generation

Researchers focus on attribute-based generations in two directions: training-based and inference-based approaches. The training-based methods either update the entire model or attach the model with additional parameters. They explore different methods, including pre-training conditional language models (Keskar et al., 2019; Zhang et al., 2020) and fine-tuning the PLM to incorporate desirable attributes (Bakker et al., 2022). Cocon (Chan et al., 2020) conditions on word- and phrase-level content to steer generation. Bakker et al. (2022) fine-tune through reinforcement learning and design a reward function for evaluating whether the generation agrees with the constraint. Besides, Qian et al. (2022) propose to learn attribute-specific prompts and Yu et al. (2021) train attribute-agnostic alignment functions. These approaches are becoming increasingly expensive due to the growing size of recent PLMs (Liu et al., 2021).

Many studies investigate inference-based strategies that affect the generation probability while freezing the PLM. PPLM (Dathathri et al., 2019) updates the hidden states toward the target tokens. GeDi (Krause et al., 2021) and FUDGE (Yang and Klein, 2021) alter the next word probability according to a step-wise attribute discriminator or bag of words. DEXPERTS (Liu et al., 2021) combines the output distributions from attribute-specific expert and anti-expert models. There are also studies that either consider attributes in energy-based models (Khalifa et al., 2020; Mireshghallah et al., 2022) or propose attribute-sensitive decoding algorithms (Kumar et al., 2021; Gu et al., 2022). Nevertheless, these studies guide the off-the-shelf PLMs implicitly with signals from other models and do not explicitly leverage retrieval systems. Therefore, as an inference-based approach, our method constructs a retrieval repository to augment attribute-based generation.

2.2 Retrieval-augmented Text Generation

Retrieval-augmented text generation assists the generative model with the information retrieval technique. It achieves state-of-the-practice results in

[1] Our code is available at github.com/araloak/grace

many tasks, including dialogue generation (Wu et al., 2021; Zhang et al., 2021), machine translation (Khandelwal et al., 2021; Meng et al., 2022), and language modeling (Khandelwal et al., 2020).

The community explores different ways to integrate the retrieved data into text generation. One line of work requires training models to learn to use retrieval knowledge. Bulte and Tezcan (2019); Xu et al. (2020) augment the model inputs by retrieving and concatenating similar samples. Hua et al. (2019); Bapna and Firat (2019); Izacard and Grave (2021) encode the retrieved texts and fuse them with attention mechanisms. Another line of studies explicitly extracts a skeleton from the retrieved data and trains the model to complete or revise it (Guu et al., 2018; Cai et al., 2019a,b). Another group is training-free methods that directly incorporate the retrieval results at the inference stage. Wang et al. (2022) prompt PLM with retrieved similar samples. Khandelwal et al. (2020); He et al. (2021); Khandelwal et al. (2021) facilitate inference with cached PLM context representations.

Our work belongs to the training-free approach. To the best of our knowledge, the existing methods do not conduct controllable retrieval in attribute-based generation, which is the target of this paper.

3 Method

Our framework consists of three parts (see Fig. 2): (1) **Attribute Discriminator** conducts attribute classification with a discriminator D to evaluate if a given context satisfies the target attribute. (2) **Retrieval Repository** builds a repository R with unlabeled corpora, which carries a mapping of a context X_n to its next word x_{n+1}. R supports reading operations to provide information that is semantically similar to the query and related to the target attribute. (3) **Generator** generates a sentence based on a prefix with a PLM G. At each step, G retrieves (read) information from R, reduces the effect of domain-specific vocabulary, and integrates it into a neural network model to generate the next word.

The above modules collaborate to conduct attribute-based generation. We design a gradient-guided retrieval-generation framework that steers generation toward the target attribute at each step and polishes the retrieved text guided by the gradient, where the gradient respects the target attribute (mentioned in Sec 3.4).

3.1 Attribute Discriminator

D consists of a context encoder, a classification layer, and a language modeling layer. The encoder transfers texts to context representations. The classification layer maps a context representation to an attribute vector, which can be used for attribute classification with an additional softmax layer. The language modeling layer maps a context representation to word probability distribution. We perform the classification with the encoder and classification layer. To obtain D, we initialize our encoder and language modeling layer with a pre-trained language model. Then, we fine-tune the encoder and the classification layer on a classification dataset.

3.2 Retrieval Repository

3.2.1 Repository Construction

We construct a retrieval repository R on unlabeled corpora via our discriminator D and generator G. The repository comprises numerous items, each containing three vectors (r^s, r^c, v^c) representing the semantics, attribute-augmented semantics, and attribute distribution of a given context.

For a sentence $X_n = \{x_1, x_2, ...x_n\}$, a subsequence is $X_i = \{x_1, x_2, ...x_i\}$ for any $i <= n$. To construct the repository R, for every subsequence X_i of every sentence in the corpora, we take the following steps: 1) G is a frozen PLM, we calculate X_i's context representation r^s by the text encoder in G. 2)We feed X_i to D's encoder to obtain its attribute-augmented context representation r^c. 3) We feed X_{i+1} to D's encoder and then the classification layer to obtain its attribute vector v^c. Finally, we define (r^s, r^c, v^c) as a repository item for X_i (see the repository items in Fig. 2). Notice that v^c measures the attribute distribution considering the next word of the current subsequence.

3.2.2 Repository Retrieval

A controllable retrieval finds the repository items similar to the query and concerning the target attribute. To retrieve for a given query text, we feed the context to the generator G to obtain a context representation r^s. Then, we search for the items whose r^s are highly similar to the query's r^s mentioned above from the repository. Further, we retrieve two sets of items with high attribute relevance as retrieval results.

$$\begin{aligned} &P_{kNN}(x_{i+1}|c, X_i) \propto \\ &P_{kNN}(x_{i+1}|X_i) * P(c|X_i, x_{i+1}) \end{aligned} \quad (1)$$

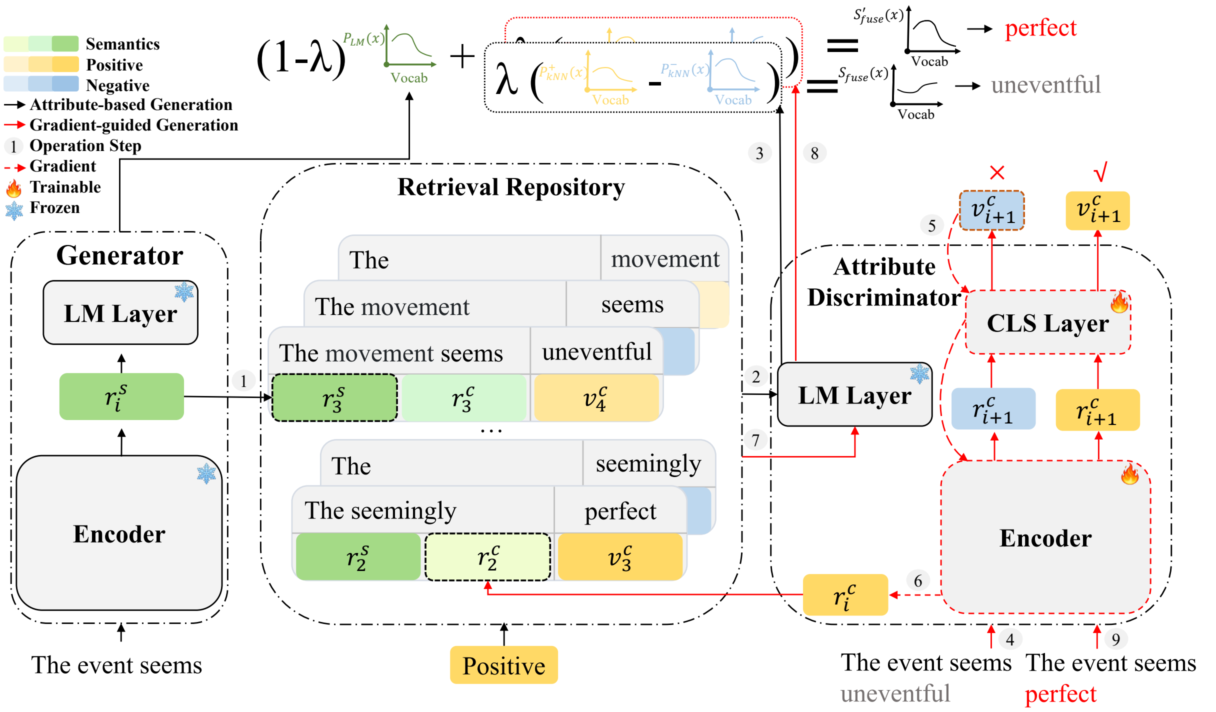

Figure 2: The architecture of GRACE. The operation steps demonstrate the i-th step of attribute-based generation with positive sentiment as the target attribute. Given the prefix "The event seems", GRACE generates "uneventful" (step 1 to 3), update the context representation for gradient-guided generation (step 4 to 8), and accepts "perfect" as the desired output (step 9).

The idea of retrieval follows the heuristic in Eq. 1 inspired by Krause et al. (2021) (see deductions in APP. A). In the retrieval results, higher $P_{kNN}(x_{i+1}|c, X_i)$ indicates (1) a better semantic coherence between the next word x_{i+1} and subsequence X_i, and (2) a higher attribute relevance with the attribute c. We design the following strategies to model the two probabilities accordingly:

- **Semantic Retrieval.** To boost $P_{kNN}(x_{i+1}|X_i)$, we search for items similar to the context X_i in semantic. As step 1 in Fig. 2, we take X_i's context representation $r^s_{X_i}$ from G as the input. We search for K nearest items from the repository according to the similarities between the stored items' context representations r^s and $r^s_{X_i}$. The algorithm returns a set $\mathcal{N}$, which provides auxiliary semantic information to facilitate the next word prediction (Khandelwal et al., 2020).

- **Attribute Retrieval.** We select two subsets of highly attribute-relevant items to increase $P(c|X_i, x_{i+1})$. First, we select items from $\mathcal{N}$ to compose a subset $\mathcal{N}^+$, where similarities between the items' attribute vectors v^c and the target attribute c excel a threshold p. The similarity is the cosine similarity between the item's attribute vectors v^c and the one-hot representation of c. v^c measures the attribute distribution considering the next word of a subsequence. Thereby, we increase the possibility of c by considering the next step generation preference. We denote $\neg c$ as the anti-target attribute [2] and obtain $\mathcal{N}^-$ following the above procedure considering $\neg c$. (The following Sec. 3.3.1 employs $\mathcal{N}^-$ to remove the domain bias from the retrieved information). In this way, we acquire items whose attributes are the most relevant to the target attribute.

Finally, the retrieval operation returns $\mathcal{N}^+$ and $\mathcal{N}^-$. The two sets contain items that are highly correlated with the target attribute and non-target attributes, respectively. Notice that the context representations in both $\mathcal{N}^+$ and $\mathcal{N}^-$ are semantically consistent with the current subsequence.

3.3 Generator

The generator G generates texts based on a given prefix and considers the target attribute. At each generation step, G retrieves from R, removes the

[2] $\neg c$ includes all the undesirable attributes. For example, if the target attribute c is "Technology" and there are four attributes in total, all the remaining attributes are $\neg c$ (i.e. Business, World News, Sports in the Agnews dataset).

irrelevant domain bias from the retrieval results, and integrates them into the generation model to produce the next token.

3.3.1 Representation Debiasing

We resolve the domain bias from the retrieved information and aim to eliminate domain-specific information from the generated sequences. We call the processing "debiasing". In most existing attribute-based generation methods, domain bias exists in the generated text where its attribute entangles with the text domain since the attribute-based training corpora usually come from a limited set of domains (Yu et al., 2021) (see App. F).

At the i-th generation step, G encodes the current subsequence to query the repository R to obtain two sets of items: $\mathcal{N}^+$ and $\mathcal{N}^-$. Afterward, we feed the attribute-augmented context representations r^c from each set into D's language modeling layer to obtain the next word probability distributions. Then, we average the values within each set and acquire $P^+_{kNN}(x_{i+1}|c, X_i)$ and $P^-_{kNN}(x_{i+1}|\neg c, X_i)$ for $\mathcal{N}^+$ and $\mathcal{N}^-$, respectively. Lastly, we calculate their difference to obtain $\Delta P(x_{i+1}|c, X_i) = P^+_{kNN}(x_{i+1}|c, X_i) - P^-_{kNN}(x_{i+1}|\neg c, X_i)$.

The intuition is that when the retrieval repository is rich in domain-specific expressions, the retrieval results of c alone may produce many domain-specific language patterns that are not necessarily relevant to the desirable attribute. However, if a word has a high probability on both $P^+_{kNN}(x_{i+1}|c, X_i)$ and $P^-_{kNN}(x_{i+1}|\neg c, X_i)$ by retrieving with both c and $\neg c$, the word is likely critical to the domain instead of the target attribute c. If $P^+_{kNN}(x_{i+1}|c, X_i)$ is high while the $P^-_{kNN}(x_{i+1}|\neg c, X_i)$ is relatively low, it indicates that the word x_{i+1} is insignificant to the domain but essential to the target attribute. Therefore, the above operation eliminates the domain bias in X_i's semantic-similar neighbors originating from R's repository corpora.

3.3.2 Representation Integration

We design a strategy to integrate the debiased information into PLM's probability to produce the next word. Intuitively, a token is desirable if it is consistent with the given context and closely related to the target attribute. We denote the word probability of the PLM in G as $P_{LM}(x_{i+1}|X_i)$ and integrate it with $\Delta P(x_{i+1}|c, X_i)$ as:

$$\begin{aligned} &S_{fuse}(x_{i+1}, c, X_i) = \\ &\lambda \Delta P(x_{i+1}|c, X_i) + (1 - \lambda) * P_{LM}(x_{i+1}|X_i) \end{aligned} \tag{2}$$

In Eq. 2, λ is a factor measuring the controllability of the target attribute c in predicting the next word. We consider $\lambda(i)$ as a step-dependent control signal that decreases linearly with the generation step i:

$$\lambda(i) = \begin{cases} \dfrac{\lambda_{min} - \lambda_0}{I} * i + \lambda_0 & i <= I \\ \lambda_{min} & i > I, \end{cases}$$

where λ_0 is the initial rate at the 0-th step, λ_{min} is the minimum rate, and I is a pre-defined step number. With fixed λ_0 and λ_{min}, a larger I allows more steps to receive higher controllability.

After the integration, we normalize the score and use the existing decoding strategy (e.g., top-k sampling) to generate the next token.

3.4 Gradient-guided Generation

We propose a gradient-guided generation to pilot the generation toward the target attribute. We iteratively evaluate the current subsequence at each step and revise it until its attribute becomes satisfactory.

- **Subsequence Evaluation.** We evaluate if the current subsequence satisfies the target attribute. We concatenate the generated word x_{i+1} at the i-th step with the current subsequence X_i to obtain X_{i+1}. Then, we feed X_{i+1} into the discriminator D and determine whether it matches the target attribute (in Sec. 3.1). We accept X_{i+1} for the next generation step if it satisfies the target attribute. Otherwise, we save the gradient of D's encoder and classification layer $\Delta\Theta$ to help update X_{i+1}.

- **Gradient-guided Subsequence Update.** We enhance the subsequence's relevance with the target attribute to help generate words with stronger attribute intensity. We optimize D's encoder and classification layer according to $\Delta\Theta$ to obtain $D_{\Theta-\Delta\Theta}$. We feed X_i into the encoder of $D_{\Theta-\Delta\Theta}$ to acquire the updated attribute-augmented context representation $r_i^{'c}$. Furthermore, based on $r_i^{'c}$, we employ the retrieval steps 3.2.2 and generation steps 3.3 introduced in the above modules to obtain new retrieval results and generate a new word x'_{i+1}. So far, we have completed an iteration of the gradient-guided generation.

When $D_{\Theta-\Delta\Theta}$ is optimized toward the target attribute, its encoder produces context representations containing richer attribute-related information, which is helpful for retrieving texts with the target attribute. Hence, r_i^c matches with items more related to the target attribute during retrieval, which in turn helps generate the next token $x_{i+1}^{'}$ more related to the desirable attribute.

In our framework, we first train the discriminator D (Sec. 3.1) and build the retrieval repository R (Sec. 3.2.1). At the i-th generation step, we follow the retrieval steps in Sec.3.2.2 and use G to generate a new word (Sec. 3.3). Afterward, gradient-guided generation (Sec. 3.4) optimizes the generation results in iterations, which calls retrieval (Sec. 3.2.2) and generation (Sec. 3.3) until it satisfies the attribute requirement. [3]

4 Experiments

4.1 Experimental Settings

Hyperparameters. We experiment on sentiment- and topic-controlled generation tasks. We initialize the D and G with GPT2-medium (Radford et al., 2019). We build our repository with FAISS (Johnson et al., 2021) for fast retrieval. To evaluate GRACE in different control intensities, we experiment on GRACE-20, GRACE-40, and GRACE-80, whose threshold step numbers I are set to 20, 40, and 80, respectively. We follow the reported settings for the baselines. More details are in App. E.

Datasets. We use one-half of the IMDB (Maas et al., 2011) dataset to train our discriminator for the sentiment-controlled generation and use another half of the IMDB, the DailyDialog (Li et al., 2017), and the Amazon (Ni et al., 2019) dataset to build the retrieval repository. Following Yu et al. (2021), We use one-half of the Agnews dataset (Zhang et al., 2015) to train a topic classifier for evaluation in the topic-controlled generation. We use another half of the Agnews dataset to train our discriminator and build the retrieval repository, which also contains the target sentences of the Xsum (Narayan et al., 2018) dataset. We follow the prefixes in Dathathri et al. (2019) to prompt the sentiment- and topic-controlled generation.

Evaluation Metrics. We follow the standard practice in attribute-based generation for automatic evaluation (Dathathri et al., 2019; Liu et al., 2021). (1) Following Liu et al. (2021), we measure *Attribute Relevance* with a HuggingFace's sentiment classifier [4] that is trained on SST-2 dataset (Socher et al., 2013) to evaluate whether the generation results satisfy its target sentiment. Following Yu et al. (2021), we train another BERT-based topic classifier with the above subset of the Agnews dataset. (2) Following (Dathathri et al., 2019), we evaluate the *fluency* of the generated text with model-based perplexity (PPL) via GPT2-large.

For human evaluation, we evaluate the generated text on overall quality (**Qual**), attribute relevance (**Attr**), and domain resemblance (**Domain**) with a 5-point rating scheme. **Qual** measures whether the generated text is grammatically correct and semantically appropriate. **Attr** evaluates whether the generation output agrees with the desirable attribute. **Domain** evaluates how likely the generation result belongs to the domain of the data that trains the discriminator[5]. For GRACE, **Domain** also evaluates whether its generation seems like the text from the repository corpora.

Baselines. **GPT2-F** concatenates attribute with the generation prefix and fine-tunes GPT2-medium. **PPLM** (Dathathri et al., 2019) perturbs a PLM's hidden states based on gradients from the discriminator or bag of words to control the generation. **FUDGE** (Yang and Klein, 2021) trains a discriminator to determine whether the future generation satisfies the target attribute. **GeDi** (Krause et al., 2021) uses GPT2-XL for generation and increases the probability of attribute-related words with Bayes' Rule. For a fair comparison, we also implement **GeDi-M** with GPT2-medium and fine-tune it on the retrieval corpora to obtain **GeDi-M-F**. **AA** (Yu et al., 2021) learns an attribute alignment to guide the PLM for attribute-based generation. Based on BERT, **MM** (Mireshghallah et al., 2022) samples attribute-related texts according to a combination of scores from off-the-shelf PLMs. Except for **GeDi** and **MM**, our baselines are based on GPT2-medium for a fair comparison.

4.2 Overall Performance

Fig. 3 and Tab. 1 show the results of all methods on automatic and human evaluations in both sentiment- and topic-controlled generation. En-

[3] The gradient update in Sec. 3.4 only affects the current generation output and does not influence other generated sentences.

[4] huggingface.co/gchhablani/bert-base-cased-finetuned-sst2

[5] **Domain** is inapplicable for methods that control with a bag of keywords (e.g., FUDGE, MM, and PPLM in the topic-controlled generation).

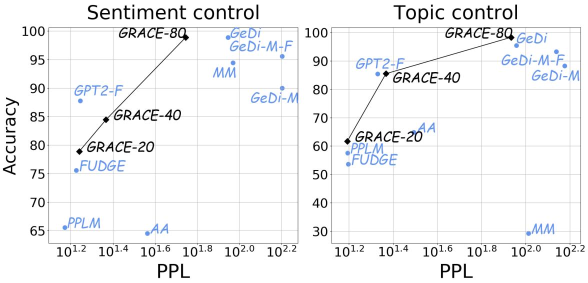

Figure 3: Overall performance of all methods on automatic evaluation. The markers in blue are the baselines. The black markers are GRACE with different retrieval steps.

	Sentiment Control			Topic Control		
Metrics	Attr ↑	Qual ↑	Domain ↓	Attr ↑	Qual ↑	Domain ↓
GPT2-F	3.74	**3.75**	3.89	4.17	**4.14**	4.18
PPLM	3.41	2.98	**1.36**	3.62	3.43	-
FUDGE	3.43	3.02	-	3.74	3.57	-
GeDi	3.96	2.78	1.38	4.32	3.54	2.28
GeDi-M	3.82	2.68	1.46	4.23	3.37	2.42
GeDi-M-F	3.93	2.66	1.39	4.13	3.44	2.68
AA	3.05	3.06	2.13	3.86	3.42	3.02
MM	3.78	2.63	-	2.76	2.34	-
GRACE-20	3.52	3.15	**1.36**	3.87	4.02	**2.27**
GRACE-40	3.62	2.83	1.58	4.03	3.86	2.36
GRACE-80	**4.08**	2.70	1.57	**4.39**	3.64	2.54

Table 1: Human evaluation of all models. The best results are in bold. Results second to the best are with underlines. Kappa score (Fleiss, 1971) among annotators is 0.59 (moderate agreement among annotators).

hancing the controllability of attributes tends to result in a less fluent generation, and vice versa (Liu et al., 2021). Therefore, we demonstrate the automatic evaluations in Fig. 3 to show that GRACE achieves a better trade-off between attribute controlling (accuracy) and generation fluency (PPL). Except for GPT2-F, GRACE achieves the best performance in both automatic and human evaluations. While GPT2-F excels GRACE in attribute accuracy, it is the worst in domain resemblance, indicating that fine-tuning on domain-specific data makes the PLM malfunction in the other domains. (see cases in Tab. 13). Existing labeled datasets for attribute-based generation only cover very few text domains (e.g., movie and restaurant reviews), thus limiting PLM's generation ability in other domains (Krause et al., 2021; Yu et al., 2021). GRACE-20 outperforms FUDGE, PPLM, and AA in attribute accuracy when GRACE-20 achieves similar or better PPL in the sentiment-controlled generation. PPLM and FUDGE behave similarly in the topic-controlled generation. Although PPLM and FUDGE achieve low PPL, their Qual is worse than GRACE-20. The reason is that PPLM may degenerate toward repeating the same word when the PLM's latent representations are not properly updated (Dathathri et al., 2019) (see cases in Tab. 10). Similarly, FUDGE may repeat specific keywords because it increases the possibilities of a limited number of keywords (see cases in Tab. 9).

In both sentiment- and topic-controlled generation, GeDi, GeDi-M, GeDi-M-F, and MM have higher PPL when their attribute accuracy is similar to or worse than GRACE-80. GeDi underperforms GRACE-80 in automatic and human evaluation even with a larger PLM. Notice that GeDi-M-F performs worse than GRACE-80, meaning that fine-tuning PLM on the retrieval corpora is suboptimal in incorporating the attribute information into generation. MM's attribute accuracy drops in the topic-controlled generation. However, it maintains a high PPL and is low on Qual, meaning that its decrease in controlling accuracy does not lead to the gain of text fluency.

By adjusting the threshold step number I, GRACE can control the trade-off between text fluency and attribute accuracy. GRACE allows near 100% attribute accuracy and can achieve a low PPL of 12.99, which is − R (equivalent to generating with GPT2 only) in Fig. 4. With the increase of retrieving steps in GRACE, its attribute accuracy improves. Notice that the accuracy improvement between retrieving 20 and 40 steps is more significant than the improvement between 40 and 80 steps. The trade-off between perplexity and attribute accuracy is more efficient during the early generation steps (i.e. GRACE-20). App. 4.6 exemplifies that GRACE excels the baselines in a case study. We report the exact PPL and attribute accuracy in App. B and show that GRACE still outperforms the baselines when we adjust their hyperparameter to re-balance the attribute accuracy and generation fluency during generation.

4.3 Ablation Study

	Sentiment Control			Topic Control		
Metrics	Attr ↑	Qual ↑	Domain ↓	Attr ↑	Qual ↑	Domain ↓
− R	2.54	3.78	1.54	1.87	4.16	2.46
− Attr R	2.86	**3.82**	1.62	2.02	**4.22**	2.38
− Semantic R	2.41	1.56	**1.25**	3.85	1.77	2.35
− Debias	3.56	2.88	4.03	3.93	3.44	4.13
− D Rep	3.61	2.76	1.61	3.86	3.75	2.63
− Revision	3.28	3.08	1.42	3.80	3.85	2.39
GRACE-20	3.52	3.15	1.36	3.87	4.02	**2.27**
GRACE-40	3.62	2.83	1.58	4.03	3.86	2.36
GRACE-80	**4.08**	2.70	1.57	**4.39**	3.64	2.54

Table 2: Human evaluation of the ablation study. The best results are in bold. Results second to the best are with underlines. Kappa score among annotators is 0.54 (moderate agreement among annotators).

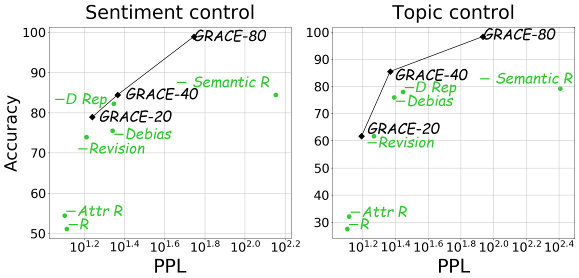

Figure 4: Ablation studies on model components. − R indicates generating without retrieval.− Attr R and − Semantic R indicate removing the attribute and semantic retrieval stage, respectively. − Debias means ignoring our debiasing method, and $-D$ Rep replace the context representation from D with that from G in representation integration. − Revision means skipping the gradient-guided generation.

Fig. 4 and Tab. 2 show the ablation studies on our model components[6]. Compared to the model variants, our full model with different retrieving steps (GRACE-20, GRACE-40, and GRACE-80) achieves the best attribute accuracy under similar perplexity. − R generates without retrieving from the repository and is equivalent to generating with PLM only. − Attr R discards the attribute retrieval stage and integrates the outputs from the semantic retrieval into text generation. − R and − Attr R perform poorly on attribute accuracy, indicating that the controllable retrieval with the attribute is crucial in controlling the generation direction (see App. C for analysis on retrieval results). − Semantic R retrieves from the repository ignoring the context representations and only considers attribute vectors. GRACE-80 outperforms − Semantic R in both metrics, showcasing that semantic retrieval helps generate fluent texts (Khandelwal et al., 2020). As − Semantic R produce unreadable texts, its domain resemblance is relatively low.$-D$ Rep integrates the context representation distilled from the frozen PLM instead of the discriminator to predict word probability. $-D$ Rep achieves a certain degree of controllability; however, the accuracy is still lower than GRACE-40. Information distilled from the attribute-agnostic PLM is less sensitive to the attribute than the that from the attribute discriminator, resulting in poor controllability. It verifies that the attribute-sensitive discriminator produces biased context representation toward attributes. − Debias does not consider the anti-target control signal while integrating word probabilities and retains domain bias in retrieval results that leads the generation toward a fixed style. − Debias is the highest in domain resemblance and underperforms GRACE-40 in attribute accuracy, rendering the effectiveness of our debiasing method. − Revision generates without the gradient-guided generation scheme to revise the poorly controlled generation and achieves low attribute accuracy. Its poor performance indicates that our gradient-guided generation is crucial to accurately steer the generation toward the target attribute.

4.4 Analysis of the Attribute-augmented Context Representation

To visualize the improvement of our variants, we show the entanglement of context representations with different attributes in Fig. 5, which analyzes the attribute information encoded in the attribute-augmented context representation. Given the same prefix under different attributes, we display the context representation r^s from the generator G, the attribute-augmented context representation r^c from the attribute discriminator D, and the updated $r^{'c}$ from the gradient-guided generation at each generation step in Fig. 5 using t-SNE. From left to right of the figure, the distribution of representations in the vector space with the same attribute becomes less sparse. Besides, the representations with different attributes are more clearly dispersed. Trained on the attribute-sensitive dataset, D encodes attribute information into r^c, making it more distinguishable in the vector space. Therefore, it encourages the generation to favor attribute-related words. r^c is further optimized toward the target attribute in the gradient-guided generation. Therefore, the updated context representation $r^{'c}$ concerning the same attribute is more concentrated, and the $r^{'c}$ with different attributes is more separable. Hence, $r^{'c}$ can match with more attribute-related items and help the gradient-guided generation to update the subsequence toward the desired direction.

4.5

sectionAnalysis of Inference Speed We analyze the time overhead of our method against other inference-based approaches. For a sentence of 80 words, GRACE requires 10 seconds per generation, while GeDi, FDUGE, and PPLM take 4, 6, and 30 seconds per generation, respectively. MM takes more than 360 seconds to generate and optimize a sentence. GRACE is slightly slower than GeDi and FDUGE but much faster than PPLM and MM.

[6] We set $I = 80$ for our model variants.

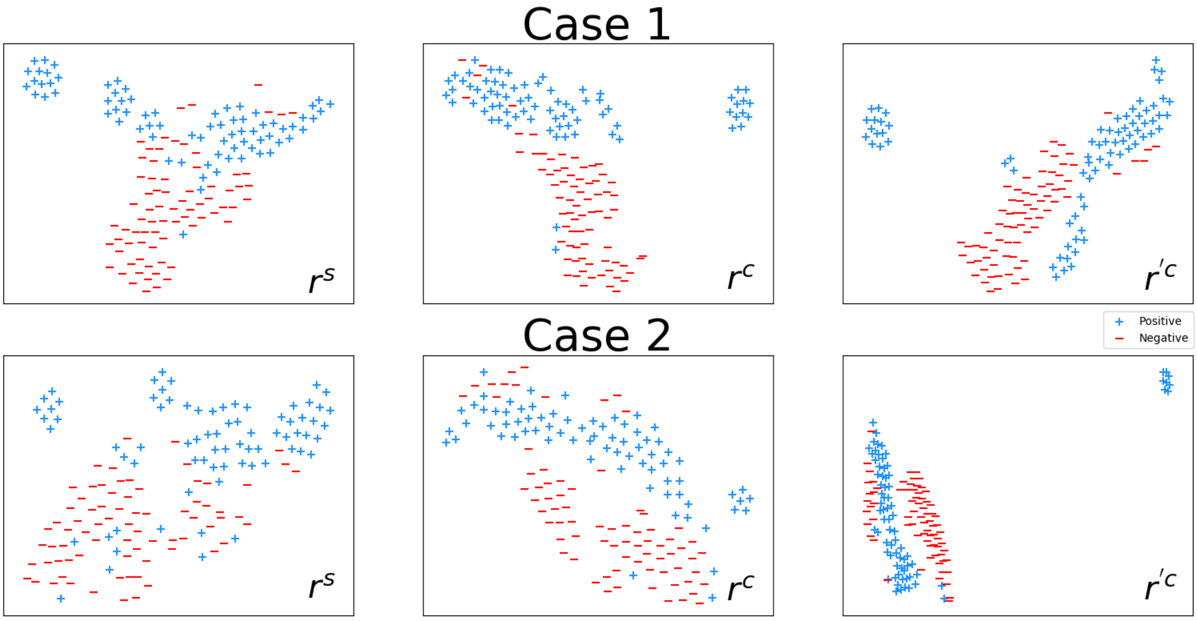

Figure 5: Visualization of different context representations from two random cases of the sentiment-controlled generation. Each marker represents a subsequence context representation during the generation.

In our method, the retrieval and gradient backpropagation are the most time-consuming operations. In the experiments, we find that the early generation sets the tone for the entire generation and is the key to achieving a controlled generation. For example, if the generation starts with "The pizza is awful", the generated result tends to imply a negative sentiment. Therefore, we provide the strongest control signal in the early stage through the step-dependent λ that declines with generation and stop retrieving after a few steps. Based on the same observation, we also limit the number of iterations of the gradient-guided generation to save more time. Our generation speed can be further reduced with other speed-up strategies and better hardware support. In the future, we will explore faster generation schemes.

4.6 Case Study

We demonstrate cases of each attribute in both sentiment- and topic-controlled generation in tables from Tab. 8 to Tab. 12. GRACE produces fluent and attribute-related sentences in all cases. PPLM sometimes degenerates when its update size is inappropriate (see Tab. 10). FUDGE increases the possibilities of the given attribute-related bag-of-words, thus tends to repeat specific keywords despite their incoherence (see Tab. 9 and Tab. 11). GeDi may generate unsatisfying sentences that are seemingly fluent but irrelevant in semantics. AA is relatively inefficient in controlling the generation toward the target attribute (see Tab. 10). MM is likely to produce less fluent sentences with grammatical mistakes. Augmented by the retrieval corpora, GRACE produces text with few repetitions and is semantically consistent among sentences.

5 Conclusion

We propose GRACE, an attribute-based generation framework that controls the generation through controllable retrieval. We train a discriminator to distinguish attributes and build a retrieval repository with unlabeled corpora. We design strategies to remove the domain bias from the retrieval information. Moreover, we propose a gradient-guided generation scheme that iteratively updates the retrieval toward higher attribute relevance. Experimental results on two attribute-based generation tasks show that GRACE outperforms strong baselines in generation quality and attribute relevance.

6 Acknowledgement

This work is supported by the following foundations: the National Natural Science Foundation of China under Grant No.62025208, the Xiangjiang Laboratory Foundation under Grant No.22XJ01012, and 2022 International Postdoctoral Exchange Fellowship Program (Talent-Introduction Program) under Grant No. YJ20220260.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Michiel A. Bakker, Martin J Chadwick, Hannah Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew Botvinick, and Christopher Summerfield. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in Neural Information Processing Systems*.

Ankur Bapna and Orhan Firat. 2019. Non-parametric adaptation for neural machine translation. In *ACL*. Association for Computational Linguistics.

Bram Bulte and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *ACL*. Association for Computational Linguistics.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019a. Skeleton-to-response: Dialogue generation guided by retrieval memory. In *ACL*, pages 1219–1228, Minneapolis, Minnesota. Association for Computational Linguistics.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. 2019b. Retrieval-guided dialogue response generation via a matching-to-generation framework. In *EMNLP-IJCNLP*, pages 1866–1875. Association for Computational Linguistics.

Alvin Chan, Yew-Soon Ong, Bill Pung, Aston Zhang, and Jie Fu. 2020. Cocon: A self-supervised approach for controlled text generation. In *ICLR*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. In *ICLR*.

Yu Duan, Canwen Xu, Jiaxin Pei, Jialong Han, and Chenliang Li. 2020. Pre-train and plug-in: Flexible conditional text generation with variational auto-encoders. In *ACL*, pages 253–262. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Jiaming Wu, Heng Gong, and Bing Qin. 2022. Improving controllable text generation with position-aware weighted decoding. In *Findings ACL 2022*, pages 3449–3467.

Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *TACL*, 6.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021. Efficient nearest neighbor language models. In *EMNLP*, pages 5703–5714. Association for Computational Linguistics.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *ACL*, pages 2661–2672. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *EACL*, pages 874–880. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

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. 2020. A distributional approach to controlled text generation. In *International Conference on Learning Representations*.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *International Conference on Learning Representations*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

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 EMNLP*, pages 4929–4952. Association for Computational Linguistics.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021. Controlled text generation as continuous optimization with multiple constraints. In *Advances in Neural Information Processing Systems*, volume 34, pages 14542–14554. Curran Associates, Inc.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*, pages 986–995. Asian Federation of Natural Language Processing.

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 *ACL*, pages 6691–6706. Association for Computational Linguistics.

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 *ACL*, pages 142–150. Association for Computational Linguistics.

Yuxian Meng, Xiaoya Li, Xiayu Zheng, Fei Wu, Xiaofei Sun, Tianwei Zhang, and Jiwei Li. 2022. Fast nearest neighbor machine translation. In *Findings of ACL*, pages 555–565. 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 *ACL*, pages 401–415. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *EMNLP*, pages 1797–1807. Association for Computational Linguistics.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *EMNLP-IJCNLP*, pages 188–197. Association for Computational Linguistics.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of EMNLP*, pages 3973–3997. Association for Computational Linguistics.

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring controllable text generation techniques. In *COLING*, pages 1–14. International Committee on Computational Linguistics.

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

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

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pages 1631–1642. Association for Computational Linguistics.

Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In *ACL*, pages 3170–3179. Association for Computational Linguistics.

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, and Bill Dolan. 2021. A controllable model of grounded response generation. In *AAAI 2021*.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *EMNLP*.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *NAACL*, pages 3511–3535. Association for Computational Linguistics.

Dian Yu, Zhou Yu, and Kenji Sagae. 2021. Attribute alignment: Controlling text generation from pre-trained language models. In *Findings of EMNLP*, pages 2251–2268. Association for Computational Linguistics.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *arXiv preprint arXiv:2201.05337*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Yizhe Zhang, Siqi Sun, Xiang Gao, Yuwei Fang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2021. Retgen: A joint framework for retrieval and grounded text generation modeling. In *AAAI 2022*.

Yizhe Zhang, Guoyin Wang, Chunyuan Li, Zhe Gan, Chris Brockett, and Bill Dolan. 2020. POINTER: Constrained progressive text generation via insertion-based generative pre-training. In *EMNLP*, pages 8649–8670. Association for Computational Linguistics.

A Deduction of the Retrieval Heuristic

By Bayes' Theorem,

$$\begin{aligned} &P(c, x_{i+1}, X_i) \\ &= P(c|x_{i+1}, X_i) * P(x_{i+1}|X_i) * P(X_i) \end{aligned} \tag{3}$$

$$\begin{aligned} &P(c, x_{i+1}, X_i) \\ &= P(x_{i+1}|c, X_i) * P(c|X_i) * P(X_i), \end{aligned} \tag{4}$$

so that we have:

$$P(x_{i+1}|c, X_i) = \frac{P(x_{i+1}|X_i) * P(c|X_i, x_{i+1})}{P(c|X_i)}. \tag{5}$$

Given the current subsequence X_i and the target attribute c, $P(c|X_i)$ is determined by the discriminator. Therefore, we obtain:

$$P(x_{i+1}|c, X_i) \propto \\ P(x_{i+1}|X_i) * P(c|X_i, x_{i+1}), \tag{6}$$

which is Eq. 1

B Details of Overall Perforamnce

We tune the hyperparameters in baseline models that affect the controllability of the target attribute to show that GRACE allows a more efficient trade-off between attribute accuracy and generation fluency. We conduct experiments on our baselines except for GPT2-F since its generation lacks a clear signal to measure the controllability of the attribute. Besides, GPT2-F generates texts like its source domain, which are inapplicable in most situations (see examples in Tab. 13). Within GeDi-based baselines, GeDi outperforms GeDi-M and GeDi-M-F in Fig. 3. Therefore, we tune GeDi to compare GRACE. As shown in Fig. 6, GRACE outperforms all baseline approaches under different settings. We show the baseline models' best performance in Tab. 3.

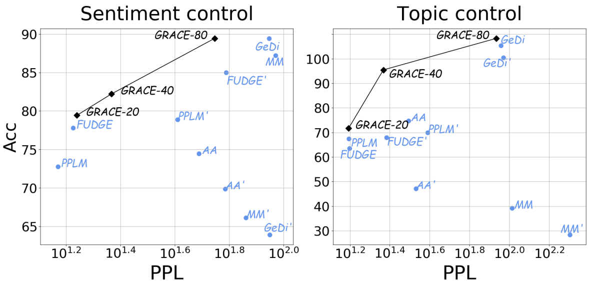

Figure 6: The details of the overall performance. The blue markers are baselines. The black markers are GRACE with different I.

	Sentiment Control		Topic Control	
	PPL	Acc	PPL	Acc
GPT2-F	17.58	87.78	21.38	85.42
PPLM	14.82	65.56	15.66	57.49
FUDGE	16.84	75.56	15.76	53.58
GeDi	88.39	98.89	90.73	95.42
GeDi-M	159.61	90.00	150.02	88.30
GeDi-M-F	159.64	95.56	137.51	93.25
AA	36.62	64.49	31.22	64.80
MM	93.48	94.44	103.12	29.17
GRACE-20	17.40	78.89	15.59	61.67
GRACE-40	23.30	84.44	23.32	85.48
GRACE-80	55.74	98.89	86.12	98.33

Table 3: The results of automatic evaluation of all models on the sentiment- and topic-controlled generation.

C Analysis of Retrieval Results

To exemplify the effectiveness of our retrieval method in providing semantically appropriate and attribute-related information, we display cases of the retrieval results with their future generations in Tab. 6 and Tab. 7. As each retrieved item comes from a piece of context X_i in the retrieval corpora, we collect the context's next word x_{i+1} in a set N-BOW. Besides, we show the top-100 high-probability words in P-BOW from the word distribution $P^{+}_{kNN}(x_{i+1}|c, X_i)$, which is interpreted from the retrieved context representations. We observe that both sets provide many semantically consistent and attribute-relevant word candidates. Besides, we find that P-BOW contains more diverse word candidates that are more intensively co-related with the desirable attribute than N-BOW. For example, in Tab. 6, many words like "NASA", "Mars", and "Pluto", are unique in P-BOW. The reason is that N-BOW collects words that appear in the retrieval corpora, while P-BOW derives from the PLM's generations, thus generalizing well to other potential preferable words. It verifies that our retrieval method can supply many coherent and attribute-relevant word candidates for each generation step.

D Multi-attribute Controlled Gereation

Although we build GRACE considering a single attribute, our method can satisfy multiple attributes by retrieving with multiple attributes. Specifically, after the Semantic Retrieval in Sec. 3.2.2, we retrieve items that satisfy the desirable attributes. Apart from this, we follow the settings in the single-attribute controlled generation to generate texts. We showcase that apply GRACE to multi-attribute controlled generation in Tab.4.

E Implementation Details

We implement GRACE based on an open-source text generation framework Fairseq [7]. We perform our experiments on a single GeForce RTX 3090 GPU with 24GB memory. To train our discriminator and the topic classifier, we set the dropout rate to 0.1 and use the Adam optimizer with a learning rate of 1e-5. The dimension of the context representations in the retrieval repository is 1024. We

[11] www.yelp.com/dataset

[7] github.com/facebookresearch/fairseq

	[Sport]
[Negative]	The country's most popular sport is a pathetic substitute for a sport for adults. It has no value in a country where the average person earns less than the minimum wage. If the average person in China had an income of US $10,000 a year, he or she could enjoy a life of leisure that would include sports such as tennis and kendo and other games in which the participants would be#
[World News]	More importantly, he said, his men will face a tougher enemy in Iraq against their own team. Japan have been fit for the first time since England's defeat in the 2004 World Cup and could play at home next week as they attempt to qualify for next summer's World Cup finals in South Africa. "We have to be very confident," said Japan forward Yasushi Nishikawa. "We have got#<|endoftext|>
[Business]	Furthermore, he has signed the club record £10.ththm deal.Head coach as well as that's the main driver for he side with £30m to play, it's the new manager who's the driving force in terms of recruitment and the new manager is the one who has the best chance to win. The only thing that will be missing from the squad will be a couple#<|endoftext|>

Table 4: The generation results of GRACE when we combine the attribute [Sports] with [Negative], [World News], and [Business], respectively. Texts with underlines are the given prefixes.

Attribute	Domain	Dataset	Statistics
Sentiment	Movie Review	IMDB (Maas et al., 2011)	50K
		SST-2 (Socher et al., 2013)	10K
	Product Review	Amazon (Ni et al., 2019)	233M
	Review	Yelp[11]	1M
Topic	News	Agnews (Zhang et al., 2015)	128K
	Wikipedia	DBpedia (Auer et al., 2007)	0.6M

Table 5: Domain of mainstream datasets for the attribute-based generation.

set the K and p in Sec. 3.2.2 as 1000 and 0.9 for retrieval and set the λ_0 and λ_{min} in Sec. 3.3.2 as 0.8 and 0.4 for representation integration. We set the maximum iteration number in Sec. 3.4 to 1. We freeze the GPT2-medium for the generation and use top-k sampling as the decoding scheme with $k = 10$. We run GRACE five times to obtain the evaluation results.

In the topic-controlled generation, PPLM and FUDGE control with a bag of topic-related words and experiment on topics that are different from GRACE. We collect keywords with a similar number on our topics and implement them to compare with our method. FUDGE does not experiment with the sentiment-controlled generation in its work. Similarly, we collect keywords representing different sentiments and follow the setting in its topic-controlled generation to guide the generation. As MM also controls generation with keywords, we implement MM with the above lists of words. Following Yang and Klein (2021), we set the maximum sentence length to 80 for all models. For each model, we run the generation 3 times on each prefix, thus obtaining 240 sentences (4 topics × 20 prefixes × 3) for the topic-controlled generation. Similarly, we obtain 90 sentences (2 sentiments × 15 prefixes × 3) for the sentiment-controlled generation.

F Discussion of Existing Attribute-sensitive Datasets

We display several labeled datasets commonly utilized for attribute-based generation tasks in Tab. 5. All these datasets imply attributes in specific text domains. Fine-tuning on these texts makes the PLM entangle the attributes with domain-specific characteristics. Thus, the generation results of fine-tuned PLM tend to be biased toward the training data domain. For example, when fine-tuned on IMDB datasets for sentiment-controlled generation, GPT2 tends to generate texts thet seem like movie reviews (see Tab. 13 for the generation results).

G Limitations

Our approach requires training a discriminator with an attribute classification dataset, which may be expensive in some scenarios. However, it is still applicable by collecting a small set of attribute-sensitive training instances and applying data augmentation techniques.

Our method is hard to achieve fine-grained control. We aim to address attribute-based generation that conditions on a given style, sentiment, toxicity, or topic. However, it cannot condition on a piece of content to control the generation. We encourage future works to explore retrieval-augmented generation with fine-grained control signals.

	[Technology]
Current Generation	In brief, you can be the best
N-BOW	the the palace through our following screen to rocket like IPv the things fuel- congrat competition, resolution present struggle shipments DJ A A company mountain- Hubble phone Computer Bas boss Fear trem Mold 14 Tomb the next model bird coordinated spacecraft the a In India's the details swimming pin carn pept viewing stellar a andised Z and provider One'ES ES Adapter shot Port the 2
P-BOW	-, the and NASA of a through computer 3 this phone is The 4 things " A Assault has mission for solar planet way video Like device Mars/ This An on letter Hubble Earth optical science but idea project (company resolution that or Not known provider an experiment in 2 planetary was probe. view hiatus image streaming implant Scientists structure Part PC Just digital X version As Space fossil to site super long kinds Chandra with article feature's its members New Pluto micro stuff: new system piece Theresoftware We Comet - It
Future Generation	In brief, you can be the best NASA engineer in the world and the solar system will have the most upDelly/minute mission ever. NASA for the first time has a space agency that does all of the heavy lifting - it just gets a little bit easier on its budget. By Michael R. Bresnahan. NASA has begun the search for a new space probe.#<|endoftext|>

Table 6: The word candidates in the retrieval stage when queried by the current generation. N-BOW contains words from the retrieval corpora. P-BOW contains words interpreted from the retrieved context representations. We highlight the keywords that imply the target attribute. The attribute here is [Technology]

	[Business]
Current Generation	This essay discusses it in depth in a US
N-BOW	firm and investors chain fourth online # officials officials industry banks based shares investment Department city and retailer Department officials Iron company maker operator an and banker banker founder city " Market history # services operators council group bank bar, (firm workers company hotel number chief # investor group steel officials firm and heart airline groups firm supplier retailer broadcaster broadcaster investor company and- firmth company shop- Department firm bank currents share shares operations could Do group ferryer firm firm firm firm firm companies pilots Securitiess investment (firm accounts stores giant store group quot chain banking Airlines income ' s s care management orders giant giant group firm- set# Stocking Labour properties and company bank published
P-BOW	- company giant and brand bank firm chain group retailer share's for officials,s store manager workers oil history. Securities President fund supermarket bailout plan insurer shop banker investment government outsourcing industry business council conglomerate income mogul discount governor employees (network payroll maker shares magazine or retail Federal private assets its pilots operator financial stock airline house Department unit in economy president Governor market funds work adviser of chains companies broadcaster sales hedge brands manufacturer investors insurance carrier stores to earnings designer firms shoppers dollar banks US ty owner bondToy P equity operations rival property
Future Generation	This essay discusses it in depth in a US retailer and the US bank that were downgraded a few the biggest in a loss' company. Diversification and a more efficient way to invest in a growing company, are key factors in the decision to downgrade their financial health. By David H. Miller - US stocks fell sharply on Tuesday but investors remain optimistic about a recovery of the tech industry. Investors#<|endoftext|>

Table 7: The word candidates in the retrieval stage when queried by the current generation. N-BOW contains words from the retrieval corpora. P-BOW contains words interpreted from the retrieved context representations. We highlight the keywords that imply the target attribute. The attribute here is [Business]

		Sentiment Control
GRACE	[Positive]	The city will also bring in an excellent new police chief with the goal of very quickly turning the city into a safer, more cohesive place," he says. "We've got to get people to understand that this is what we are going to do." Mayor John Tory says he's confident this year's campaign will bring in a lot more people to the streets in the fall. (#<|endoftext|>
	[Negative]	The city's police commissioner, James C. Raley, said Thursday that he just learned about the death of his friend who was killed in the line of duty, but said he did not believe his department had done enough to ensure that its officers followed department protocols when responding to calls of officers involved in serious accidents. "When I saw this tragedy unfold, I couldn't imagine a more tragic#<|endoftext|>
PPLM	[Positive]	The city of Austin is seeking a new zoning for a proposed mixed-use complex that's expected to include a large public garden and a restaurant on one corner of an existing public park.City Council passed the zoning change last month after public hearings on the
	[Negative]	The city of Houston, Texas has been hit with $500,000 in damage to its water system. The damage was caused by a fire that started in a hose. The fire caused the fire hose that is used to pump the city's
FUDGE	[Positive]	The city and its suburbs are blessed with ample natural beauty and a great variety of recreational opportunities and recreation activities; the beauty, natural beauty, and recreation opportunities are all part of what makes the area one of the best places in the country to live for recreation, recreation, recreation and recreation. I have always believed that the best way to enjoy and live in the area is to embrace the opportunity
	[Negative]	The city of Birmingham, she said, "is the worst place for women" and "is the worst for people with mental illness" — even "because it is a white, Christian community" where "there are no black people" and "neither black nor white are welcome." She said she had "fought" for her "blessing" by being "faulted
GeDi	[Positive]	The city on a hilltop will nurture and empower young girls in all ages by providing a safe place to grow into strong girls who understand that they are unique and can make or break the success of their neighborhoods, careers, schools and families.<|endoftext|>
	[Negative]	The city council said it would investigate the complaint but sent no response. IIT-Madras society lecturer Moununot Feridunhas slammed the university, claiming students were left wondering if their dreams simply never came true. Launching a legal action against it on its website, Feridunhas also warned that state-of-the washing caused drinking water levels in<|endoftext|>
AA	[Positive]	The city's old-world charm, its old-world sense of humor and its laid-back way of playing things cool may turn some people off, but it's a great movie for people who like their romances to have that french realism.'– kurt wimmer.'s the greatest date movie in years!' – ellen pompeo.'– michel pic
	[Negative]	The city plays too little." These sorts of conversations are not uncommon. But their effect on very small features of the cultural fabric of the city— the buildings themselves, and the people living inside them—can still be profoundly damaging to those plans, beyond just the suggestion that they could be altered for the better. Take, for example, Kelly David Herman's Colusns Wilderness, a major piece of city infrastructure
MM	[Positive]	The city is most often characterized by high - quality sunny blue skies (the shape of an egg) , attractive inhabitants (often women) , and numerous characters both american and european , including robin hood ; its naturally lynding nature (and thereby its large immigrant population) ; peppers and tomatoes , (particularly those eaten by donald trump) ; entertainment , dining and recreation ; health and well - being ;
	[Negative]	The city was surrounded by thick gray clouds . overfilled families - perhaps even dozens - were huddled under the faltering clouds , while the rest - not just the girls , charlotte , madeline , madeline and sutton - clung desperately to the meaning of the message , but to charlotte , to me - to god god help us all - and to the mistaken identity of thayer and nearly all the others .

Table 8: Examples of the generated sentences of different baseline approaches under the control of different sentiments. The texts with underlines are the given generation prefixes. We highlight the words that are highly related to the target sentiment. The sentiments are [Positive] and [Negative].

	Topic Control: [Technology]
GRACE	This essay discusses the planetary climate in order to show how the global climate system is changing and how the changes could affect us. In the process we will see how climate can be used to understand the evolution of life in the universe and how our own evolution might have evolved. We will see how climate is changing because of the actions that have taken place in the past and what this means in terms of the#<|endoftext|>
	The issue focused on researchers at Stanford, Harvard and MIT who had created an experimental method of analyzing images of animals and their brains for patterns of electrical activity. The researchers had shown that animals with abnormal electrical activity in their brains displayed patterns of activity in the cortex, a part of the brain that controls language, movement and other complex behaviors. The pattern of electrical activity was then analyzed to determine which brain areas#<|endoftext|>
PPLM	This essay discusses developments in technology and technology policy, technology, innovation, technology transfer, and the role of technology in promoting peace and stability in the twenty-first century. It focuses specifically on the role of technology in supporting peace and stability through technology. It draws on the extensive literature on innovation as a tool of development in the United Nations, especially the work of scholars such as Richard Feynman and
	The issue focused on a new technology: the technology that allows people to communicate in a way that makes sense for their particular situation."We've been developing technology that allows people to communicate through the Internet through their phones, through their laptops, through their computers, through social media and through their computers, without having to have any sort of human interaction at all with the person they're communicating with
FUDGE	This essay discusses the relationship of gender, politics, and the media and technology industries. It is not about technology and technology-based technology systems.The technology industry is one of the most important technological technologies in history because of its potential to transform the way we live. Technology is technology, the technology industry is technology, and the technology industry is technology. It is technology and technology, technology and technology, technology
	The issue focused on the use of the new technology in the industry of virtual reality technology, which is currently being developed by the technology company VR Technologies.The issue focused on the use of the new technology in the industry of virtual technology, which is currently being developed by the technology firm VR Technologies.The debate over technology technologyThe discussion on technology in the technology industry was sparked by the controversy surrounding
GeDi	This essay discusses Unix philosophy in depth. It recreates Unix in its most accessible form, offering original commentary on some of Unix's most influential program formulation and implementation principles or "methods" (think sophisticated real-world unix development such as Harper's List or H-UML). This edition of the book, written between April 1998 and January 1999 but originally published by Elisabeth<|endoftext|>
	The issue focused on Linux fragmentation, which experts point out as one of the biggest problems with proprietary content distribution platforms such as Ubuntu. While Apple has not been a regular user of Linux since it officially announced its support in 2010, both changes come after years of criticism that Google and other companies are largely helping developers build fan projects aimed at using Chrome's technology. Debian is emerging from such controversy<|endoftext|>
AA	This essay discusses two aspects of Apple Computer #39;s most recent hardware update. The first issue concerns the use of the XPC software technology. The second concerns making use of new capabilities built into Windows 2000, Internet Explorer and the like. Apple Computer has since acknowledged the use of an exploit in its OS/2 Personal Computer... it claims the exploit does #39;s not appear to
	The issue focused on technology for connecting RFID (radio frequency identification) systems with payments. Sony bought IBM and it began selling RFID cards. Its technology integrates IBM's operating system and integrated RFID reader. The card connects to the reader and scans a QR code. IBM version 4 of the IBM iQuote card prints on an RFID chip onto a paper label. IBM iQuote cards sell for between $129
MM	This essay discusses " human impulses " . knots (essays) [240pp .] 1908 : knots and other essays on electrical engineering and technology , published by the pratt institute press . hali [original leaf print .] 1911 : [william henry mccook] and charles darwin [original leaf print .] on an american treadmill , war - weary sailors exchange respect and friendship .
	The issue focused on cliches ; economic hardships ; genre - bending ideas ; a new cover design - bursting at the seams - incorporating bowie himself ' s photographs (including an exclusive interview with bandmate andrew lloyd webber) ; ugly and righteous head : ugly and righteous head songs (which were meditations on suffering and degradation) , with each verse both before and after returning to its original ;

Table 9: Examples of the generated sentences of different baseline approaches under the control of different topics. The texts with underlines are the given generation prefixes. We highlight the words that are highly related to the target topic. The topic here is [Technology].

	Topic Control: [Sports]
GRACE	Prior to this Sunday's game with the Saints, the Saints had a record of 2-3-1 and had a 3:5 lead at halftime. In the second half, the Saints led 14-10, but they were down 21-13 to the New Orleans Saints, 27-27 in the third quarter. With the Saints trailing 24-24 and having the ball at#<|endoftext|>
	In this essay at least, we are not dealing with the players of the game, nor the managers of the club, but rather with the players themselves. We are not trying to prove anything, but to show the fact that there is not much that can be learned from players' performance statistics. We want our readers to be able to make an informed decision about their own football. If you#<|endoftext|>
PPLM	Prior to this year, I had no idea the term 'poster' even existed.A poster is a type of poster with a printed or digital design that is attached to a vehicle.I have no idea what it was like to get my driver's licence, nor what it is like for the public to watch a sports event on television on my television.In my sport, there
	In this essay we look at the most influential women in football and discuss their achievements on and off the field."The game will not have to change in the NFL if NFL owners are not willing to make the league football football," NFL commissioner Roger Goodell said Tuesday during an interview on NFL Network's NFL Football pregame show. Goodell also said he believes football football football football football football football football football football NFL football
FUDGE	Prior to this season, many athletes, athletes, athletes, sports fans, sports fans and sports fans were complaining about the "fitness gap" between white athletes and blacks. The fitness gap was created by many of the following factors:- Athletes of color are often less experienced- Athletes of color are often more likely to get injured- Athletes of color have
	In this essay, I'll discuss the three main players of the sport: the sport's elite athletes, the amateur athletes, and the professional athletes. These three groups share a common goal: achieving Olympic medals in the sport they compete in.The elite athletes In the sports of gymnastics, judo, soccer, swimming, track, and field, professional athletes are the only Olympic
GeDi	Prior to this season, fans would sit out games or boycott the club if they felt every player was booing raucous music everyone should know about. Now players come together in a wait-for-hormones atmosphere where playing loud tunes doesn't send an offensive message because the constant pressure made showing up excites them. "You gotta keep everything positive as we're just men<|endoftext|>
	In this essay sports writers will summarize and analyze every game played during each week (ends May 24th 2013). Want to win a key enemy mission? No problem. Won't want three more hours of mindless galactic War Games? Good luck! KeyGame analysis umbrella concept assumes these post games affect 5 points total based on results out of each team. Regrets: The Gold Medal Implied Team stats<|endoftext|>
AA	Prior to this past weekend's N.H.L. hockey game in Buffalo, fans of the New England team had a chance to see some of the younger players make an appearance. And boy, did they show some of that athleticism. Playing with the young guns was a nice break from what #39;s been going on all season. The regular season is a nice break from the... er... madness.
	In this essay, I will analyze a public exploit in a lab environment, see the alerts generated by an intrusion detection system, and then do some packet analysis of the malicious binary in order to better understand it.As I understand it, this binary is a variant of the Shell Insert Bot (SHB) variant, which is used in... hellip; many malicious virus attacks today.
MM	Prior to this, the betrayer was collected by yuri petrushkin , hardcover , 1986 . (hardcover [ ] ; publisher : muller - raythen , germany , folio ; publisher : muller - raythen , germany , hardcover , 1988) rabbinate samson , rabbinate of salonica and about a hundred others : a comparative study (1720 - 1740) , e . g .
	In this essay, friedlander describes two separate but related deaf and mute races , named mutants , and x - men (or simply martians) . behavior with regard to mutants varies from jealousy or affection (love) to hostility . phaethonus and marihuana look like brightly colored apes , while those he considers to be mutants also look like martians , ending up looking brusque .

Table 10: Examples of the generated sentences of different baseline approaches under the control of different topics. The texts with underlines are the given generation prefixes. We highlight the words that are highly related to the target topic. The topic here is [Sports].

	Topic Control: [Business]
GRACE	To review, the company cut its dividend, and the stock plunged from an all time high in June. The stock was up about 2%. "It was an extraordinary period of market action," says David S. Daley, director of research at Wedbush Securities. "I think this is the best-selling company at a time where the S&P 500 is in its worst stretch in#<|endoftext|>
	More importantly, a households survey shows that $2 trillion in household savings has been lost since 2009 due to financial collapse. It's time to take a hard look at the current economic reality in the US and demand more reforms from Washington. In a recent op-ed for Forbes.com, the CEO of Goldman Sachs, Lloyd Blankfein, argues that the US needs to focus on the#<|endoftext|>
PPLM	To review, this is one of the first companies to launch a smartphone with NFC technology, and its product is a smart business that can make financial services companies profitable businesses. The company offers businesses a business-to-business solution with an online banking service, and it can also offer financial services companies the financial services business services industry, which can make companies companies companies companies companies companies companies companies companies companies companies companies companies companies
	More importantly, if the companies that make the products you buy are able to pay their suppliers to use their products, the companies can pay off customers, too.In other words, you could see the companies paying their suppliers to do something. That's what's happened at several companies, from companies like Facebook to companies like Google.The companies are paying suppliers to use their products. This has created
FUDGE	To review, there are many companies that are trying to provide an easy-to-use experience for the consumer, and there are a few companies in the market that provide this functionality. In addition, there's no single company that provides the best experience to the consumer in order to maintain the quality of their product.A great company to consider in this regard is Microsoft. They provide a great
	More importantly, companies have a responsibility to ensure the best of their products are being used in the most efficient and cost-efficient ways by their workers. This is especially true when the company's workforce includes many foreign workers, as the U.S. government has recently acknowledged. The U.S. Trade Representative (USTR) is currently working with the U.S. Trade Administration and other stakeholders
GeDi	To review, Dollar Tree agreed to repurchase $4.2 billion dollar bonds, Orange Grocery agreed to complete appropriate write downs ca $2 billion dollar bond and Pipe Life repealing significant loan carrybacks the Dollar Tree Cash Stocks fell $900 dollars due to past finance mistakes was sold under advisement.. year. Check for updated information today as trading data are updated Company trades month Invest in Gram<|endoftext|>
	More importantly, investors should understand that capital inflows by FRBNY securities currently account for almost all of the continued rally in domestic home prices." Emily Category of The Wall Street Journal and Candy Chen contributed to this article.<|endoftext|>
AA	To review, there is one thing that should be avoided in this model; acquisitions. There are too many GMs over at Toys ""R Us"" and other retailers that have the idea that acquisitions can help drive higher profits. This is exactly what has happened in the toy business during the past few years, and it has significantly exaggerated growth. As my colleague the late Robert Bloch said: ""The acquisition shows that
	More importantly, each HP vendor has taken a second look at the risks and challenges involved in bringing 1,000 series servers to market, with a focus on ensuring that the operating systems supporting these servers meet the same security standards as vproducts shipped on the mainframes before 2005. The vendor's annual certification programs will also take a second look at how the required enhancements work. HPC vendors will be reviewing security policy, distribution
MM	To review, finance , distribution and marketing , confessions of a sentimental fool closed its doors . under goodspeed and company , columbia records (cbs) commissioned crosby , stills and nash (" wonderland " / " enough wildness for an angel ") , elaine paige (" my fifth studio album , wonderland ") and tom & jerry (" pastime and pleasure ") to produce six additional albums .
	More importantly, in size and in shape , samson was charmed by his surroundings . he and dixon had been best friends forever , and still there was hope for him here in arcadia . dixon had been responsible for samson . he had actually been the caretaker . additionally , he had also been the financial advisor to the j . farrell company (which meant the property was open to further development) until now .

Table 11: Examples of the generated sentences of different baseline approaches under the control of different topics. The texts with underlines are the given generation prefixes. We highlight the words that are highly related to the target topic. The topic here is [Business].

	Topic Control: [World News]
GRACE	The connection up to the Palestinian Authority's security apparatus to the West Bank is particularly sensitive. Israeli security services, for instance, are believed to operate in the territories as well as in Israel. It would seem that the Palestinians would like it otherwise, but they are not the only ones who do not trust the Palestinian Authority security services or their agents. Israel has a long history of using the PA#<|endoftext|>
	The relationship between the Foreign Service and the intelligence community is one of deep concern for the Obama administration. But in a sign of the new reality facing intelligence professionals, the Obama administration is also considering a new proposal that would require them to register with the government and report on any foreign contacts they may have made with foreign agents — in other words, a new way to keep tabs on potential foreign threats. #<|endoftext|>
PPLM	The connection was made in the first quarter of 2017.A senior Indian telecom minister on Friday said that India is looking into a proposal to create a national broadband network (NBN) for India today. In an interview to news agency PTI, Minister of State for Information and Broadcasting Manish Tewar today said the Indian telecom regulator has been informed of this proposal. He also said the
	The relationship between the United States and the world's media has been a tumultuous one over the past year. In the wake of the Trump presidential campaign, it has become increasingly clear that the press is being controlled. This has included a concerted effort by outlets to discredit and attack each other over the media, including Fox News and its reporting on the Trump campaign. This is in addition to the mainstream media's
FUDGE	The connection is not just for media outlets and websites. There is also a third group who have access to the data, including news outlets and bloggers, as well as academics and journalists.A report this summer from the Pew Research Center found that the Internet and social media were having a major role in spreading stories about Russia.The report found that the Internet has helped to expose the news
	The relationship is also at the heart of news coverage and public debate over the issue.The report, written by the International Council for Science.The report was released by the Council on Foreign Relations in a report entitled, The Future of Science and the World. The report's author, a senior official at the council's think tank, was quoted as saying, "It is time we
GeDi	The connection of terrorism with Sharia Law is nothing more than a parental form of anti-American propaganda. Apart from the fact that Islam declares up-dates and thus emulates State laws, this perversion without objections evolved as a propagandistic tool to distract from laika, which was specifically appropriated by Saddam Hussein's regime not during the 1991 Gulf War any longer. Sismah conducts no SC<|endoftext|>
	The relationship Iran must be led by Daddy's wishes. Mér-Ali Khomeini, the supreme leader of Iran and descendant of Mehmed Shah Pahlavi – the founder of Mullah Omar's legacy – declared you live '36 years' and 'earned your money'. A decade later he decreed that women will not take part in political or military posts or social activities unless<|endoftext|>
AA	The connection between West Germany and the United States is increasing, but the visit of President Bush will not be without controversy. Joachim Gauck visited Berlin last week with business and political leaders from the United States and Canada and met his British counterpart...We are up against more than simply a Cold War... Europe and the United States are facing very different threats. It is important to remember that each party is fully
	The relationship between nuclear and nuclear power has taken a turn for the better, with construction of new reactors and clean energy projects proliferating in many areas. Now it's heating up again, thanks to the discovery of a new heavy-water reactor beneath an ancient lake in southeastern Wisconsin that may power the #39;s main cities for another hundred years. #39; Plus: Montana regulators mull importing energy from
MM	The connection is powerful and intense but not long lasting . you demand an explanation , you laugh , or you cry . partial or complete disclosure , in turn , may avert an emotional timbre . " ignazio , why not make recent media reports available , despite your having never read news , without admitting that you may have experienced angst in the dark , brightly lit car an older woman drove ?
	The relationship aspect of the episode was criticized as " excessively emotional " . stefanie zus - suzanne zus (stella adler) an interviewer reflects on her belief that suzanne should either enjoy or not be enjoying any part of her life : who is this person ? a woman who breaks a vow made by her husband (jack mcclain) on an emotional level because of breaking her own vow ?

Table 12: Examples of the generated sentences of different baseline approaches under the control of different topics. The texts with underlines are the given generation prefixes. We highlight the words that are highly related to the target topic. The topic here is [World News].

	Training Data Domain: Movie Reviews
[Positive]	The lake scene in "The Big One", which is a real treat. It’s like the first "The Big Lebowski", but much better in every way. The story of how an American man and a Japanese man, living on opposite sides of the world, are reunited is very touching and very funny. The two actors are very talented and I am very#<|endoftext|>
	The horse is a great actor and actor is one of the best parts. The rest is not to bad of a film. This movie is one of my favorites for the story, the characters, the actors and the plot. The story was good and the horse was really good. But I have to say that the film was a big let down when I had to#<|endoftext|>
	The pizza delivery guy is the only one that can save you from the zombies, if not his brother.If the movie is good it’ll be a good film. If not it’s a waste of time and money. I’ve seen many bad zombie flicks and I’m glad I didn’t have to spend the money to watch it.The acting is decent at#<|endoftext|>
	The potato is an interesting concept in that it doesn’t require a great deal of effort to figure out how to grow them. The movie has a good idea of what to expect from this potato and its uses but there are a good number of technical difficulties that make growing potatoes difficult. I’ve seen some documentaries about potatoes but they tend to gloss over what the potato actually#<|endoftext|>
[Negative]	The lake scene in the beginning of this film was just awful. It is just a bad movie with bad acting, bad direction and bad acting by the actors! I think they should have cut this movie because it is just too much. The lake scene should have been shot underwater instead of on the beach in the first place! I really like this film, but it#<|endoftext|>
	The horse in this movie looks like a very small one, it is a horse that looks like it could easily have been put into a toy box, it does not move at all. The movie is really boring, it is not very funny at all, the acting is horrible. The horse looks more like it was made by a 4 year old, and is a#<|endoftext|>
	The pizza scene is just a horrible, awful joke. There is no substance in it and it has zero plot. It’s all just bad acting and a bad script. There are many scenes where the actors look like they are about to faint. I don’t understand why they bothered making this movie at all.They should have just stuck with a movie about a guy#<|endoftext|>
	The potato chips in my mouth were really good, but the story is so predictable that it really hurts. The movie is so predictable that if the characters are not in a relationship, it is hard to care for any of them. It has the worst acting I have ever seen. The main character who is supposed to be a lesbian and is a gay man in the#<|endoftext|>

Table 13: The randomly sampled generation results of GPT2-F on sentiment-controlled generation. The texts with underlines are the given generation prefixes. The texts in blue indicate the domain of GPT2-F’s training corpus.

ACL 2023 Responsible NLP Checklist

A For every submission:

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

☒ A2. Did you discuss any potential risks of your work?
The ethical impact of our research is the same as other text generation papers, whose ethical impact is widely discussed.

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

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

B ☑ Did you use or create scientific artifacts?

3

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

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

☑ 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)?
E

☑ 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?
E

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

☑ 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.
E

C ☑ Did you run computational experiments?

4

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

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?
E

☑ 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?
E

☑ 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.)?
E

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
4

☑ 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.?
E

☑ 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)?
E

□ 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?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.