Few-Shot Style-Conditioned LLM Text Generation via Latent Interpolation

Moemen Gaafar, Anurag Sarkar, Matthew Guzdial
Department of Computing Science
University of Alberta
{mkgaafar, sarkar.an, guzdial}@ualberta.ca

Abstract

We propose a novel, model-agnostic approach for adapting large language models (LLMs) in a few-shot manner to generate text conditioned on samples from a given author to represent that author's voice or style. Rather than predefined features, our method defines style in terms of LLM model weights and uses a variational autoencoder to construct a latent space of these weights, enabling a generic style representation. Our approach interpolates in this latent embedding space of model weights to generate novel finetuned models for low-resource authors. We evaluate this approach compared to reported results, finetuning, and prompting across three datasets. Results indicate that our method outperforms our baselines in low-resource settings.

1 Introduction

Text generation is a core NLP task that enables applications such as machine translation, summarization and question-answering (Li et al., 2022). Guiding this task such that it outputs text satisfying a certain set of constraints is referred to as controllable text generation (Zhang et al., 2023). One such form of control is style-conditioned text generation i.e., constraining the generated text such that it follows a certain writing style, usually corresponding to a specific author (Mou and Vechtomova, 2020). When generating text using large language models (LLMs), conditioning on a certain style is particularly challenging due to the prohibitive training data requirements of LLMs (Zhang et al., 2023). Successfully conditioning the style of text generated via LLMs using just a few samples holds potential for new applications such as empowering users to produce text matching their own writing voice or approximate the voice of another writer.

Existing methods for controlling text generated via LLMs include prompting, finetuning and postprocessing (Zhang et al., 2023). Among these, prompting is the least computationally demanding, but prior work has shown that it cannot reliably infer style from a few samples (Patel et al., 2022; Liu et al., 2024). On the other hand, finetuning, while theoretically capable of adapting a model to arbitrary styles, typically requires a sizeable corpus, even when combined with techniques such as low-rank adaptation (LoRA) (Hu et al., 2021). Finally, postprocessing methods which modulate the output probabilities of an LLM can condition the style of generated text in a few-shot manner (Khan et al., 2023). However even this reduced training corpus might be burdensome in applications where a user produces text in real-time. Additionally, many of these methods assume the existence of predefined style features (Lagutina et al., 2019).

In this work, we propose a novel model-agnostic approach for performing few-shot adaptation of an LLM to generate text conditioned on a target style. Our approach differs from prior methods by representing style in terms of model weights rather than predefined features and by using a variational autoencoder (VAE) to construct a latent space encoding differences in model weights, which we refer to as *weight deltas* and extract using LoRA. Thus, our approach does not require predefined style features and leverages the VAE to learn a space of possible finetuned models. This means that we represent style as not just word choice, but in terms of how different authors would respond with different meanings to the same input. This may be considered closer to representing an author's voice than stylizing content to match an author's word choice. We argue that this enables our approach to be more adaptable in terms of style representation. Given a small number of samples from an author, we generate new weight deltas by performing interpolation in our VAE latent space. We identify two major contributions: (1) a novel model-agnostic method for performing few-shot style-conditioned text generation and (2) a new approach for directly representing text style in terms of model weights.

2 Related Work

2.1 Style-Conditioned Text Generation

Style-conditioned text generation refers to the task of generating text in the style of a target author (Tikhonov and Yamshchikov, 2018). This differs from the popular NLP task of text style transfer in that the goal is not to preserve content while transferring style and thus does not require disentangling style from content. We follow Subramanian et al. (2018), who demonstrated that this disentanglement, besides being challenging, is unnecessary to model style. The standard approach for this task is to train a language model on a corpus of text in the target style (Tikhonov and Yamshchikov, 2018). However, for transformer-based LLMs, the data requirements for this task have become increasingly prohibitive (Zhang et al., 2023). Our work addresses this limitation as we aim to model arbitrary text styles in a few-shot setting while leveraging the generation capabilities of modern LLMs.

LLMs have shown strong capabilities in multiple NLP tasks in both zero-shot and few-shot settings (Zhang et al., 2023). STYLL (Patel et al., 2022) demonstrates that LLMs are able to perform style transfer with arbitrary styles via prompting. However, this approach prompts the LLM to classify the target style using specific attributes which are later used to perform style transfer, similar to work by Reif et al. (2022). This approach performed satisfactorily for attribute-based style transfer but did not demonstrate the ability to model complex authorial styles. Liu et al. (2024) address these issues by proposing ASTRAPOP, an RL-based actor-critic approach to style transfer. However, while this does well on medium-sized corpora, its performance is inconclusive on smaller corpora belonging to a single author. Instead of prompting, Subramani et al. (2022) extract latent vectors from a pretrained LLM which produce desired target sentences when added to the model's hidden states. However, obtaining vectors corresponding to a specific style requires separately extracting vectors for each sentence in that style. Similarly to our own work, Jin et al. (2024) make use of LoRA-based weight increments associated with particular style features, though we do not predefine style features. Additionally, Ilharco et al. (2022) also adapt model behavior using task-specific weights obtained via finetuning, though they focus on editing pretrained models via arithmetic operations on these task-specific vectors, whereas we generate new models via interpolation in a learned latent space of finetuned weights, motivated by prior work (Sarkar et al., 2020) demonstrating that interpolation in a learned latent space can produce meaningful style combinations in accordance with the interpolation parameters Finally, Khan et al. (2023) proposed StyleMC, a unified approach to style transfer and stylized text generation using future discriminators (Yang and Klein, 2021) that leverage pretrained style embeddings. By instead using model weights directly, we forego the need for any predefined notions of style. Defining style in terms of model weights follows prior work (Guzdial and Riedl, 2019) that has demonstrated the validity of recombining weights from different models for approximating new styles. We compare our approach to StyleMC due to its published results on known datasets. We note that in all of this prior work, even those focusing on few-shot settings, assume access to sixteen or more samples. This data requirement can at times exceed hundreds of samples, proving a burden on end users.

2.2 Text Style Transfer

Similar to but distinct from style-conditioned text generation, text style transfer (TST) is the task of converting text from its source style to a target style while preserving its content (Fu et al., 2017; Jin et al., 2022). TST can be categorized into two broad types—attribute-based and authorial. Attribute-based methods aim to transform text along one or more explicitly defined stylistic dimensions (Subramanian et al., 2018; Subramani et al., 2022). Such approaches are limited due to their reliance on labeled data and their inability to model complex styles. Conversely, authorial style transfer (Jhamtani et al., 2017; Syed et al., 2019) aims to transform text to a style that is not straightforward to define explicitly, such as styles attributed to unique authors. TinyStyler (Horvitz et al., 2024) is a recent work that does this in a few-shot manner. Overall, while similar to style-conditioned generation, TST differs in seeking to preserve meaning across source and target texts whereas style-conditioned generation considers how different authors may respond to the same input. We focus on style-conditioned generation in this work and leave considerations of TST as potential future extensions to the current approach.

2.3 Low-Rank Adaptation (LoRA)

Low-Rank Adaptation (LoRA) (Hu et al., 2021) falls under a class of finetuning approaches referred

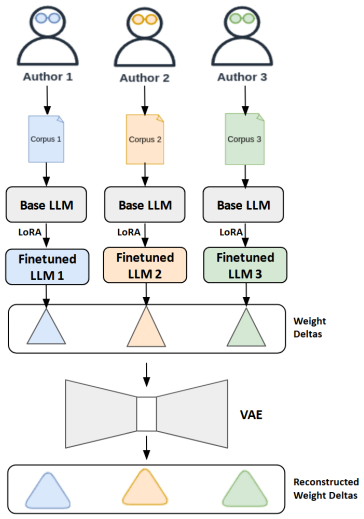


Figure 1: Overview of our approach for learning a latent space of finetuned weight deltas.

to as parameter efficient fine-tuning (PEFT) methods (Houlsby et al., 2019) which freeze the original weights of a given pretrained model and instead add a smaller set of trainable task-specific weights termed *adapters* to certain layers of the model. In LoRA, this takes the form of passing a given d-dimensional input vector x simultaneously to the frozen pretrained layer with weights W and to a learnable projection matrix A which maps x into an r-dimensional space where $r << d$. A second learnable matrix B then maps the input back to the d-dimensional space and this output is then summed with the output of the frozen layer to produce the output fed to the next layer. We rely on LoRA to make the learned space of weight deltas compact enough to afford interpolation.

3 Methodology

In this section, we present our proposed approach for few-shot style-conditioned text generation. We frame this as the task of obtaining an LLM finetuned on a given small corpus (< 10 samples) of text in a certain style. The finetuned LLM is then capable of generating text in that style. Given a pretrained autoregressive LLM $P_\Phi(y_t|x, y_{<t})$, where Φ represents the base model weights, the same LLM finetuned on a corpus C is represented as $P_{\Phi+\Delta\Phi_C}(y_t|x, y_{<t})$ where $\Delta\Phi_C$ stands for the difference in model weights (*weight delta*) on account of fine-tuning the base LLM. As a shorthand, we represent the weight deltas as Δ_C and the corresponding finetuned LLM as P_{Δ_C} in the rest of the paper. Thus, our task can be expressed as follows: given a small text corpus C^* from a certain author, we find Δ_{C^*} that, when applied to the base LLM weights, produces the finetuned LLM $P_{\Delta_{C^*}}$, which generates text that mimics the style of that author. Our overall approach consists of two steps: 1) constructing a latent space of LLM weight deltas and 2) approximating novel finetuned models via interpolation in this latent space. We discuss these steps in the following sections.

3.1 Learning a Latent Space of Weight Deltas

In order to construct a latent space of LLM weight deltas, we extract the deltas and use them to train a VAE. This process is summarized in Figure 1. We start with text corpora belonging to distinct authors exhibiting a developer-chosen number of distinct writing styles. We finetune a base LLM on each corpus using LoRA to obtain our finetuned models. The weight deltas are then extracted from these models and used as inputs to the VAE, which in turn yields a latent space of these deltas.

3.1.1 Extracting Weight Deltas via LoRA

For the pretrained LLM $P_\Phi(y_t|x, y_{<t})$, we extract a collection of model weight deltas $\Delta_1, \Delta_2, \Delta_3, \ldots, \Delta_n$ corresponding to instances of the original LLM finetuned on text corpora $C_1, C_2, C_3, \ldots, C_n$, with each corpus belonging to a distinct author. Each corpus must be large enough for finetuning to capture the style of each text, though we note this limitation does not extend to novel styles encountered after training the latent space. Thus, we can apply the approach to authors with far fewer samples than what would normally be needed to finetune the LLM. When applied to the base LLM, the extracted weight deltas produce the finetuned LLMs $P_{\Delta_1}, P_{\Delta_2}, P_{\Delta_3}, \ldots, P_{\Delta_n}$ respectively. This approach makes no assumptions about the finetuning process and thus any finetuning method could be applied so long as the differences are captured sufficiently in all or a subset of model weights. For our experiments, each corpus C_i contained at most 300 text samples, with each having a maximum of 60 tokens. In order to make the behavior of the system more predictable, we perform finetuning using the same hyperparameters for all authors so that differences between weight deltas are only dependent on the training corpora. To reduce the dimensionality of these weight deltas and the cost of finetuning, we capture approximations of the weight deltas by applying

LoRA and further reduce the dimensionality via Principal Component Analysis (PCA).

3.1.2 Embedding Weight Deltas via VAE

Using these extracted weight deltas as input, we train a variational autoencoder (VAE) (Kingma and Welling, 2013) to learn a low-dimensional latent representation of the extracted deltas. VAEs have been widely used for generating new content via interpolation and are relatively easy to train and less compute-intensive compared to other latent variable architectures. Given an input weight delta Δ_C, the VAE encoder outputs two latent vectors corresponding to the mean μ and covariance σ matrices that define a Gaussian distribution. The VAE decoder samples from this distribution and outputs the reconstructed weight delta Δ'_C. The VAE loss function is given by:

$$\mathscr{L} = \|\Delta_C - \Delta'_C\|^2 + D_{KL}(N(\mu_{\Delta_C}, \sigma_{\Delta_C}), N(0,1)),$$

In our experiments, we found it useful to weigh the KL term using a tunable parameter β (Higgins et al., 2017) giving us the loss function:

$$\mathscr{L} = \|\Delta_C - \Delta'_C\|^2 + \beta \cdot D_{KL}(N(\mu_{\Delta_C}, \sigma_{\Delta_C}), N(0,1))$$

We found that setting $\beta < 1$ helped the VAE adequately reconstruct the weight deltas due to their small scale and high variability. However, a side effect of weighing the KL term was that it led to a partially disconnected latent space made up of clusters of training samples separated by regions of random noise. As we need a space of continuous, valid weights for interpolation, we mitigated this by filtering out disconnected data points to extract a continuous latent space subset.Our filtering algorithm starts by sorting the data points in ascending order of the sum of their reconstruction and KL losses. This allows differentiating between points that the VAE fit well (low total loss) from those that it did not (high total loss). For each dataset or split of a dataset, we extract n data points with the lowest total loss ($n = 5$ for this paper's experiments), which we then average to find their centroid in the latent space. Then, we calculate the maximum of their Euclidean distance from this centroid. For each data point in the dataset, we find the distance between itself and the centroid of its dataset and compare that to the aforementioned maximum Euclidean distance multiplied by a factor. If the distance to the centroid is less than that threshold, it is considered a part of the continuous subset of the latent space. If not, it is discarded. We found that a factor of 3 enabled adequately distinguishing between points that did and did not belong to the continuous subspace of the latent space for all datasets, verified by inspecting the latent space graphically and using Euclidean distance. These results are shown in appendix sections A.3 and A.4. For the results in the main paper, we only make use of the continuous, filtered latent space.

3.2 Generating Weight Deltas via Interpolation

After training the VAE and filtering the latent space, we generate new weight deltas and their associated models via interpolation within the VAE latent space. We perform linear interpolation guided by a random sample from the finetuned models $P_{\Delta_1}, P_{\Delta_2}, P_{\Delta_3}, \ldots, P_{\Delta_n}$. This approach does not rely on any additional hyperparameters or domain knowledge, making it generalizable to other domains and datasets. To simplify interpolation, this method assumes that the topology of the latent space is smooth, which we ensure with our clustering-based filter algorithm. At test time, given a small corpus C^* containing a few text samples from an unseen author, we pick K models at random from the collection of finetuned models and perform one pass of finetuning, i.e., one-step of backprop, on each using C^*. Due to the small size of C^*, this step is not computationally intensive. We perform a single step of backprop for each group of samples to reduce the time and resources required to perform inference as well as to avoid overfitting. We found that as we increase the number of text samples and average these changes, our approximation improved. This can be seen as equivalent to doing multiple gradient steps. We compare against one-step and not multiple steps of finetuning on the same samples for a fair comparison. This one-step finetuning yields $\Delta^*_1, \Delta^*_2, \ldots, \Delta^*_K$ which vary only slightly from the original $\Delta_1, \Delta_2, \ldots, \Delta_K$ (and are thus incapable of modeling the style of C^*). We then forward $\Delta_1, \Delta_2, ..., \Delta_K$ as well as $\Delta^*_1, \Delta^*_2, ..., \Delta^*_K$ through the VAE encoder to get $\mu_1, \mu_2, ..., \mu_K$ and $\mu^*_1, \mu^*_2, ..., \mu^*_K$ respectively. We treat each pair (μ_t, μ^*_t) as a vector in the latent space moving from the original model μ_t in the direction $\vec{r}_t = \mu^*_t - \mu_t$. This gives us K latent space vectors to guide us towards a point that approximates the target model corresponding to C^* when passed through the decoder. This target model would consist of the

weight deltas representing the style of corpus C^*, as derived by the following interpolation process. For each pair of N-dimensional lines $(\vec{\mu_a}, \vec{r_a})$ and $(\vec{\mu_b}, \vec{r_b})$, assuming they point towards our target model, we want to find their intersection point in the latent space. Expressed mathematically, we find t_a and t_b that satisfy the following equation:

$$\vec{\mu_a} + t_a\vec{r_a} = \vec{\mu_b} + t_b\vec{r_b} \quad (1)$$

which can be re-written as:

$$\vec{r_a}t_a - \vec{r_b}t_b = \vec{\mu_b} - \vec{\mu_a} \quad (2)$$

or in matrix form $Ax = B$ where

$$A = \begin{bmatrix} \vec{r_a} & -\vec{r_b} \end{bmatrix}^T$$

$$B = \begin{bmatrix} \vec{\mu_b} - \vec{\mu_a} \end{bmatrix}$$

$$x = \begin{bmatrix} t_a \\ t_b \end{bmatrix}$$

Solving this equation returns the vector $\vec{\mu_{C^*}} = \vec{\mu_a} + t_a\vec{r_a} = \vec{\mu_b} + t_b\vec{r_b}$ which, when passed through the decoder, reconstructs our target Δ_{C^*}. However, since there is no guarantee that two N-dimensional lines will intersect, we modify this equation to instead find the least-square approximation:

$$\arg\min_{x \in \mathbb{R}^N} \|Ax = B\|^2, \quad (3)$$

Solving this equation returns t_a and t_b that represent the closest point on each line to the other one. In this case, we return the midpoint of the line connecting the two points as our target $\vec{\mu_{C^*}}$, which is then forwarded through the decoder to be reconstructed. For certain lines, this point falls in the negative direction of the vector from the source to the target models, indicating that the two models are diverging in one or more latent dimensions. While mathematically valid, we do not consider the interpolated models to be valid in such cases as they point to the opposite direction of the intended target. For our experiments, we sampled $K = 3$ models to perform interpolation. Interpolated models differ based on the choice of source models. For our results, for each test data point, we choose base models from a different dataset or dataset split than that which the test point belongs to. Results when testing with base models from the same dataset as the test point are shown in section A.5.3 of the appendix. We experimented with two interpolation methods that differ in how successive text samples are handled. For simple interpolation, we perform linear interpolation directly using only the changes in latent space produced in the current timestep. For accumulative interpolation, we accumulate and average the changes in the latent space produced across all previous steps until the current one. This helps stabilize the interpolated models. This stability issue is discussed in appendix section A.5.4. The results using accumulative interpolation are discussed in the results section while those obtained using simple interpolation can be found in appendix section A.6.

4 Experiments

4.1 Datasets

We evaluated our approach using three datasets based on 1) Twitter, 2) Project Gutenberg and 3) Reddit. The Twitter dataset is a subset of the Sentiment140 dataset containing tweets tagged with the Twitter handles of their authors (Go, 2009). We filtered out authors with fewer than 200 tweets to ensure there were sufficient text samples to finetune the LLM, ending up with 17 authors in total. We collected the Gutenberg dataset from the Project Gutenberg website, which offers free access to electronic books. We retrieved the top 100 most popular books of all time and treated each book as a separate author with a distinct style. We feel this is a reasonable assumption since an author's style may change from one book to another, especially in fiction. For Reddit data, we used a subset of the Reddit Million User Dataset (MUD) (Baumgartner et al., 2020). Similar to Khan et al. (2023), we focused on four subreddits with distinct styles, namely: r/wallstreetbets, r/news, r/AskHistorians and r/australia. We filtered out authors with fewer than 200 posts and picked 30 at random from each subreddit, ending up with a total of 120 authors. For each dataset, we set aside 10% of the authors as the test data and used the remaining for training. We split the datasets by author to ensure that the styles of the authors in the test data were not seen by the VAE during training. For both training set and test set authors, we further split each author's text corpus into train, validation and test sets which we used to finetune and evaluate the LLMs.

4.2 Finetuning Llama-2 with LoRAs

We used the autoregressive LLM Llama-2-7b (Touvron et al., 2023) as the base LLM in this work. Since the model has 7 billion weights, instead of finetuning the original parameters, we apply LoRA by adding an adaptation layer to each Q and V atten-

tion layer of the base LLM. LoRA enables saving on computational resources and lets us directly use the adapter weights as the weight deltas for our system. Finetuning is done via next-token prediction on each corpus using cross-entropy loss. Llama-2-7b contains 32 decoder units, each with four 4096×4096 attention layers, amounting to more than 2 billion total weights. Applying LoRA to the Q and V matrices, each unit now contains four r×4096 vectors instead where r refers to the rank of the LoRA adapters. We set r to 2 to decrease the number of weights to $\sim$1 million. Thus, for each corpus, we finetune an instance of Llama-2 using LoRA to obtain a 32×4×2×4096 weight delta. We used an alpha value of 8 and a dropout rate of 0.1. When the LoRA adapter layers are populated by the weight delta, we obtain a finetuned model that generates text in the style of that author.

4.3 VAE Training

We show the VAE architecture in the Appendix in Figure 8. We used fully-connected networks as there is no inherent structure to the weight deltas. We used a latent dimensionality of 8 and trained the VAE for 600 epochs at a learning rate of 1e-4, using the Adam optimizer and a β value of 0.03. Even though we set the LoRA rank to the smallest possible value for effective finetuning, we found the weight deltas' dimensionality to still be a bottleneck on the GPU memory during training. We thus compressed the deltas using Principal Component Analysis (PCA) due to its simplicity and negligible GPU processing requirements. We trained 32 PCA models (one for each Llama-2 decoder unit) to reduce the dimensionality of the weight deltas from 4×2×4096 to 4×2×P_d where P_d is the number of output components of the PCA model. We found PCA to be capable of reconstructing the model weight deltas, even unseen ones. For the Reddit data, setting $P_d = 400$ led to PCA explaining 98% of the average variance. The 2% loss corresponded to negligible effects on the performance of reconstructed weight deltas as verified by assessing their generated text and a quantitative assessment of their cross-entropy losses (shown in appendix A.2). We trained the VAE using these PCA-reduced deltas.

4.4 Compute Resources

Experiments were conducted using cloud resources provided by `Anonymized`, consisting of 18 CPU cores and 2xNVIDIA v100l GPUs with 32GB of memory each.

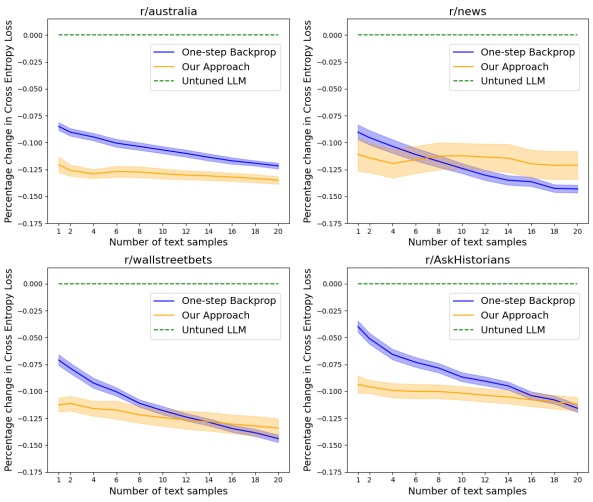


Figure 2: Cross-Entropy Loss for the Reddit dataset.

4.5 Metrics

For our experiments, we used the cross-entropy loss on the test split of the corpus of a given unseen author to compare the performance of the interpolated models to the finetuned source models. This quantitatively estimates the extent to which the interpolated models use the same words in the same order as the reference author corpus. Since this requires access to the LLM's output probabilities, it cannot be used for comparisons with closed-source models. Since LLMs have a different base cross-entropy loss for different styles, to aggregate the losses among different authors, we normalized each author's cross-entropy loss using their base LLM cross-entropy loss. We also used the Content-Independent Style Representations (CISR) (Wegmann et al., 2022) and the Universal Authorship Representation (UAR) model (Rivera-Soto et al., 2021) to generate author style embeddings and measure the cosine similarity between the test split and the generated text to approximate style similarity.

4.6 Baselines

We compared the interpolated models to the finetuned source models as well as prompting both GPT-3.5 and the base Llama-2 for style-conditioned text generation. We also compared the performance on the Reddit dataset to StyleMC which, to our knowledge, is the only prior system that targets the problem of few-shot style-conditioned text generation with a number of samples roughly equivalent to our work, albeit by using a more computationally intensive inference procedure than our approach. We directly quote the results as reported by the StyleMC authors.

Method	2 samples	16 samples
Our Approach	**0.555**	0.549
One-step Backprop	0.532	0.546
Prompting GPT-3.5	0.385	0.485
Prompting Llama-2	0.371	0.501
StyleMC	-	**0.762**†

Table 1: CISR similarity on Reddit data for different methods. † indicates value taken directly from source.

5 Results

In this section, we present the results of our experiments, organized by dataset.

5.1 Reddit Results

Figure 2 shows the cross-entropy losses obtained for the Reddit dataset, grouped by subreddit. We show the percentage change in the loss, relative to the base cross-entropy loss, on the y-axis. All other losses are below this base loss since finetuned models are better at modeling the text corpus compared to the base model due to similarities among Reddit posts even if they are taken from different subreddits. Results show that our approach outperforms finetuning for all subreddits for low numbers of text samples (< 10). As more text samples become available, the performance of our approach approximately converges to that of finetuning.

Figure 3 shows the cosine similarity scores using CISR embeddings. We see that performance is mostly equivalent between the finetuned models and our approach. Figure 4 shows results of comparing our approach with GPT-3.5 prompted to perform style-conditioned text generation. Though GPT-3.5 outperforms our approach in some cases, the variance in its performance is much larger and suggests that it is less reliable for this task. GPT-3.5 is also many times larger than our model and may contain the Reddit data in its training corpus.

Table 1 shows the CISR scores for our approach and each baseline, aggregated across all subreddits. We also include results obtained by prompting an instruction-tuned Llama-2 for comparison. Our approach outperforms all baselines for both 2 and 16 samples, besides StyleMC. The reported performance of StyleMC with 16 samples exceeds that of our approach, however, as noted before, text generation using StyleMC involves several steps of inference using a sampling process, thus requiring more computation than our approach. They also report a higher mean score for GPT-3.5 (0.56) compared to what we obtained (0.485). This may

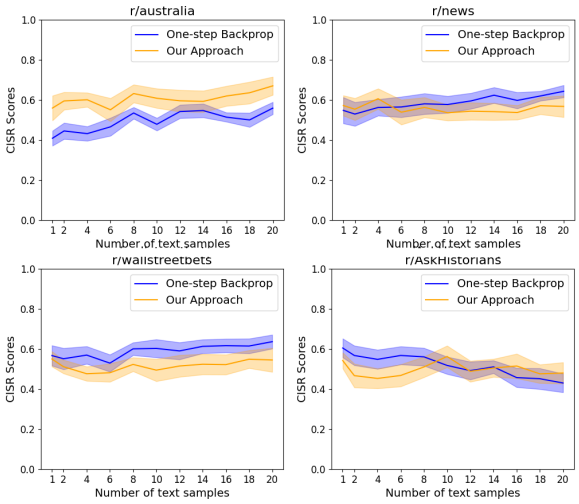


Figure 3: CISR scores for the Reddit subreddits comparing our approach with finetuning.

suggest that though we used Reddit data from the same subreddits, our choice of users might differ from those used in evaluating StyleMC.

Figure 5 shows the cosine similarities using UAR. The only subreddit which exhibits a difference between the finetuned models and our approach is *r/wallstreetbets* with performance being mostly equivalent for other subreddits. Figure 6 shows our approach compared with GPT-3.5 using UAR. Table 2 shows UAR scores for our approach and each baseline, aggregated across all subreddits. We again note a disparity between the GPT-3.5 score (0.742) reported by StyleMC compared to our results (0.649), likely due to the same reason as before. Though prompting outperforms our approach in terms of UAR scores as the number of samples increases, our approach has a smaller margin of error and is thus arguably more reliable for this task. We ran the Mann-Whitney U Test between these distributions and found our approach has a significantly higher distribution (better scores) compared to one-step backprop and prompting GPT-3.5 using 2 samples. We also note that the relative lack of change in model quality when going from 2 to 16 samples is consistent with the existing literature on weight blending/interpolation approaches (Singamsetti et al., 2021).

Our relative performance is stronger using CISR compared to UAR likely because CISR represents style independent of content, which aligns better with our task of style-conditioned text generation where unlike style transfer, content preservation is not required and could interfere with the ability of newly generated text to adhere to the desired style.

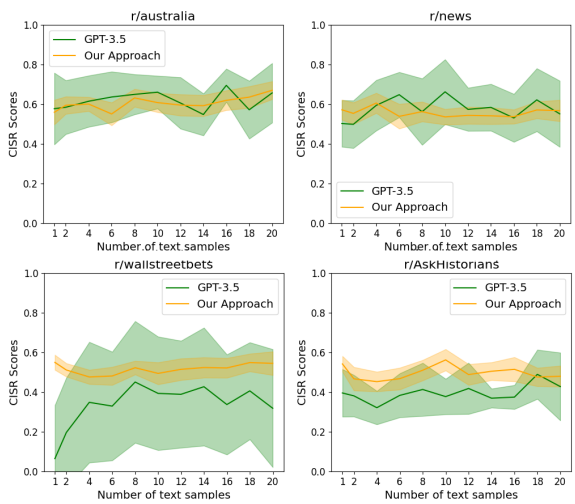


Figure 4: CISR Scores for the Reddit dataset comparing our approach with GPT-3.5.

Method	2 samples	16 samples
Our Approach	**0.609**	0.608
One-step Backprop	0.588	0.601
Prompting GPT-3.5	0.581	0.649
Prompting Llama-2	0.592	0.633
StyleMC	-	**0.849**†

Table 2: UAR similarity on Reddit data for different methods. † indicates value taken directly from source.

5.2 Twitter and Gutenberg Results

Figure 7 shows the results for the Twitter and Gutenberg datasets with the percentage change in cross-entropy loss on the y-axis. The results for Gutenberg are similar to those for Reddit with our approach outperforming finetuning for low numbers of samples and converging to finetuning as the number is increased. However, for Twitter, finetuning slightly outperforms our approach, particularly as the number of text samples is increased. This may be due to the Twitter dataset's smaller size compared to Gutenberg. Thus the VAE is unable to learn representations that are salient enough for useful interpolation. We also computed UAR and CISR scores for the Gutenberg and Twitter data but these metrics failed as the representations are trained on Reddit and are thus inadequate for representing out-of-distribution text styles. We report these results in appendix section A.8. Sample text generations are shown in appendix section A.10.

6 Conclusion

We presented an approach for style-conditioned text generation using LLMs in a few-shot setting.

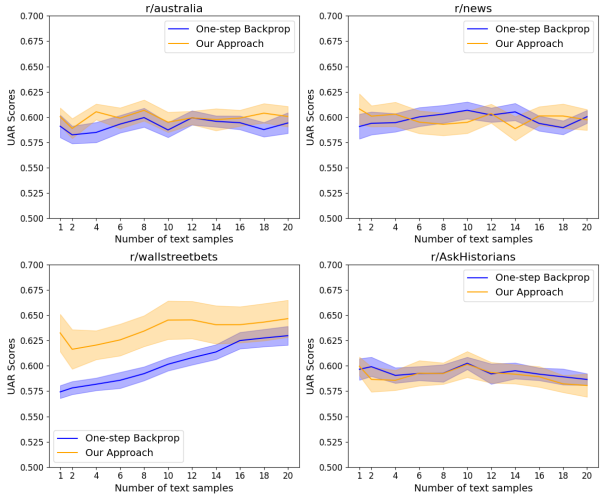


Figure 5: UAR Scores for the Reddit dataset comparing our approach with finetuning.

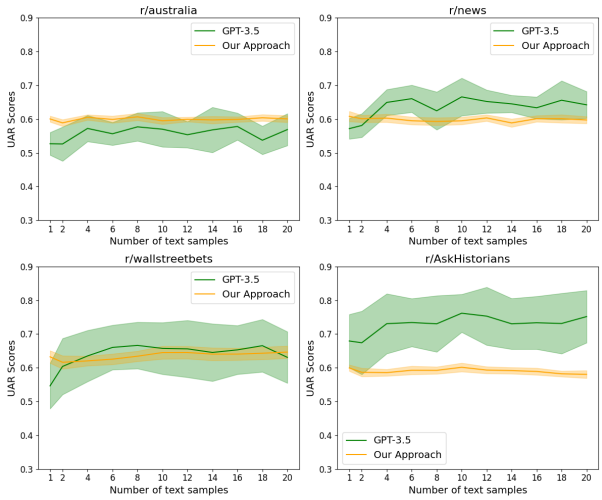


Figure 6: UAR Scores for the Reddit dataset comparing our approach with GPT-3.5.

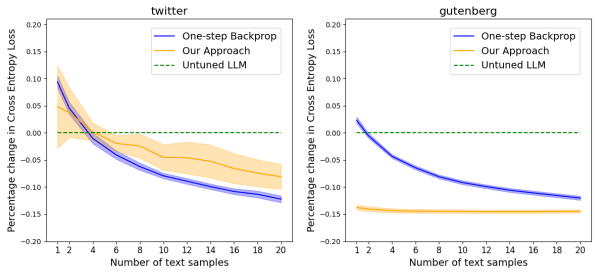


Figure 7: Cross-Entropy for Twitter and Gutenberg.

We demonstrated that a VAE is able to encode meaningful style information in the form of latent embeddings of LLM weight deltas finetuned via LoRA. We also presented evidence that interpolation in this latent space enables generating new model weights that outperform finetuning and other baselines in low-resource ($n < 10$) settings.

7 Potential Risks

We acknowledge the potential harm of dual use, or use by bad actors, from this research. Few-shot style-conditioned LLM text generation expands the possible applications of LLM technologies. But these possible applications include bad actors using this to mimic a target's writing style for impersonation or conditioning an LLM on their own style of writing to mask their use of LLM generated text. We plan to acknowledge this risk in the public release of the code but do identify that both of these applications are already possible, and more work is needed in detecting and combating such uses.

Limitations

Our current work has a number of limitations. In terms of our system itself, the inclusion of a filtering step is not ideal. For simplicity, we would prefer if we could regularize the latent space of the VAE to make filtering no longer necessary. Additionally, we used a relatively straightforward interpolation method as we wanted to test our methodology using a simple approach and thus plan on testing more complex interpolation strategies for future work.

In terms of our evaluation, we identify two major limitations. First, we only ran our full evaluation on finetuned Llama models. While we present some evidence of generality in terms of our GPT-2 results, we would ideally demonstrate stronger proof of generality by applying our approach to larger and more diverse models. We focused on a single model due to resource and time constraints but do plan to investigate this direction in future work. Second, our evaluation setup itself had a limitation in terms of only making use of proxies of human judgment. Ideally, we would be able to demonstrate human subject study results supporting our approach's ability to match a target style. While our approach is setup to allow for real-time generation, we did not have the capacity to run a human subject study. Further, we felt that it would be better to gather evidence for the potential of the approach prior to running a human subject study. However we plan to run one in the future.

References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Michael Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2017. Style transfer in text: Exploration and evaluation. *ArXiv*, abs/1711.06861.

Alec Go. 2009. Twitter sentiment classification using distant supervision.

Matthew Guzdial and Mark O Riedl. 2019. Combinets: Creativity via recombination of neural networks. In *International Conference on Computational Creativity*.

Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3.

Zachary Horvitz, Ajay Patel, Kanishk Singh, Chris Callison-Burch, Kathleen McKeown, and Zhou Yu. 2024. Tinystyler: Efficient few-shot text style transfer with authorship embeddings. *arXiv preprint arXiv:2406.15586*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

Harsh Jhamtani, Varun Gangal, Eduard H. Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. *ArXiv*, abs/1707.01161.

Chunzhen Jin, Eliot Huang, Heng Chang, Yaqi Wang, Peng Cao, and Osmar Zaiane. 2024. Reusing transferable weight increments for low-resource style generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Aleem Khan, Andrew Wang, Sophia Hager, and Nicholas Andrews. 2023. Learning to generate text in arbitrary writing styles. *ArXiv*, abs/2312.17242.

D.P. Kingma and M. Welling. 2013. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*.

Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and P.G. Demidov. 2019. A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 184–195.

Junyi Li, Tianyi Wang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2022. Pretrained language models for text generation: A survey. *Preprint*, arXiv:2201.05273.

Shuai Liu, Shantanu Agarwal, and Jonathan May. 2024. Authorship style transfer with policy optimization. *ArXiv*, abs/2403.08043.

Lili Mou and Olga Vechtomova. 2020. Stylized text generation: Approaches and applications. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 19–22, Online. Association for Computational Linguistics.

Ajay Patel, Nicholas Andrews, and Chris Callison-Burch. 2022. Low-resource authorship style transfer: Can non-famous authors be imitated?

Project Gutenberg. Project Gutenberg. `http://www.gutenberg.org`. Retrieved July 20, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anurag Sarkar, Zhihan Yang, and Seth Cooper. 2020. Conditional level generation and game blending. In *EXAG Workshop*.

Mohan Singamsetti, Anmol Mahajan, and Matthew Guzdial. 2021. Conceptual expansion neural architecture search (cenas). *arXiv preprint arXiv:2110.03144*.

Nishant Subramani, Nivedita Suresh, and Matthew E Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581.

Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text style transfer. *ArXiv*, abs/1811.00552.

Bakhtiyar Syed, Gaurav Verma, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Vasudeva Varma. 2019. Adapting language models for non-parallel author-stylized rewriting. *ArXiv*, abs/1909.09962.

Alexey Tikhonov and Ivan P. Yamshchikov. 2018. Guess who? multilingual approach for the automated generation of author-stylized poetry. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 787–794.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268.

Kevin Yang and Dan Klein. 2021. Fudge: Controlled text generation with future discriminators. *ArXiv*, abs/2104.05218.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3).

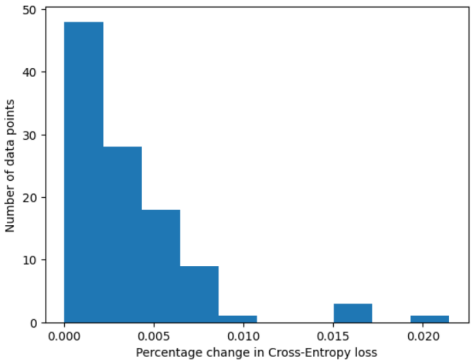
A Appendix

A.1 VAE Architecture

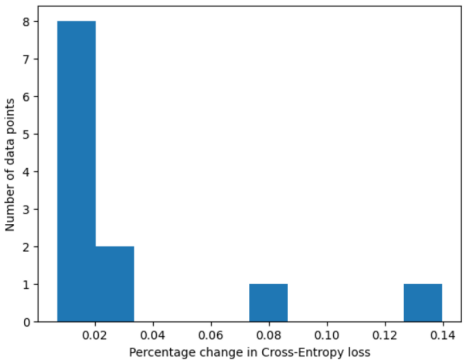
The architecture for the VAE used in this work is shown in Figure 8.

A.2 PCA-reduced Weight Deltas

As discussed in section 4.3, the difference in the percentage change in cross-entropy loss for the original and PCA reconstructed weight deltas is shown in Figure 9.



(a)



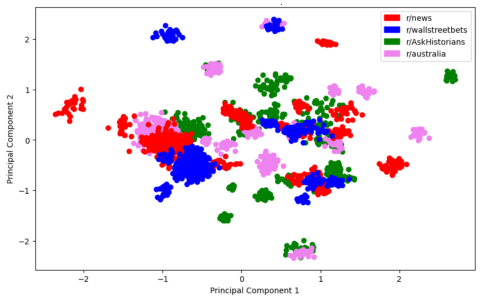
(b)

Figure 9: Histograms of the difference in percentage change in cross-entropy loss between the original and the PCA reconstructed weight deltas for (a) the train split and (b) the validation split of the Reddit dataset.

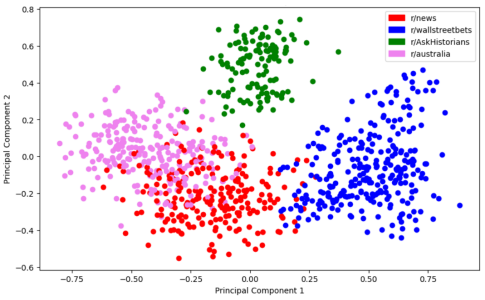
A.3 Reddit Dataset Latent Space

The latent space of the VAE trained on the Reddit dataset is shown in Figure 10a. As can be seen, the latent space is fragmented and discontinuous. However, we extracted a continuous subset of the latent space via the filtering algorithm as described in section 3.1.2, consisting of 40 data points, constituting 37% of the training set. This filtered subset of the latent space is shown in Figure 10b. We see here that the VAE is able to differentiate between the weight deltas that belong to the different subreddits without any prior information.

We confirm this observation by computing the Euclidean distance in the latent space between data points based on the subreddits they belong to. The heatmap in Figure 11b confirms that in this filtered latent subset, data points from the same subreddit are closer to each other than to data points from other subreddits. This is in contrast to the heatmap in Figure 11a which shows that this does not apply to the full, unfiltered latent space.



(a)



(b)

Figure 10: Two-dimensional projection of the VAE latent space trained on Reddit data for (a) the full latent space, and (b) the connected subset of the latent space.

A.4 Twitter and Gutenberg Latent Space

As with the Reddit data, we found that the full latent space of the VAE trained on the combined Twitter and Gutenberg dataset was disconnected. However, as shown in Figure 12, in this case, a larger portion of the data fit within the continuous subset of the latent space. 11 points and 66 points were encoded in this continuous subset for Twitter and Gutenberg respectively, accounting for 78% and 81% of their respective training data. We hypothesize that the higher percentages are due to higher inner similarity among data points within these datasets, compared to Reddit. Again, the VAE is able to discern the differences between both datasets from the weight deltas alone.

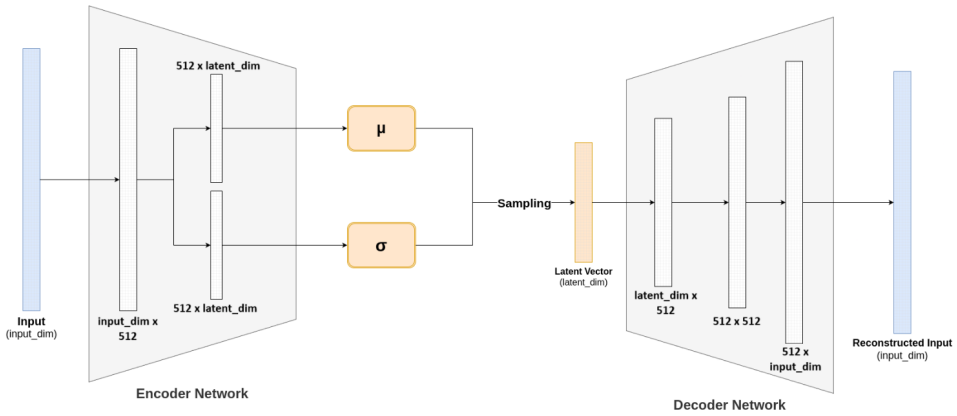
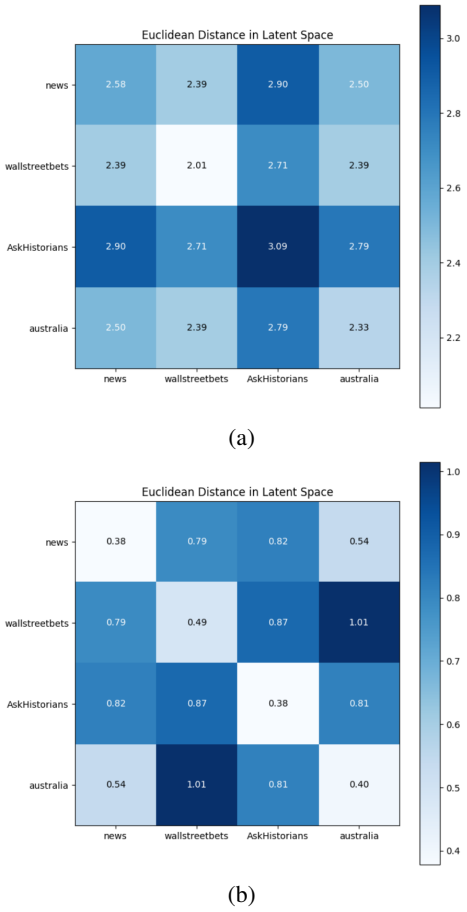


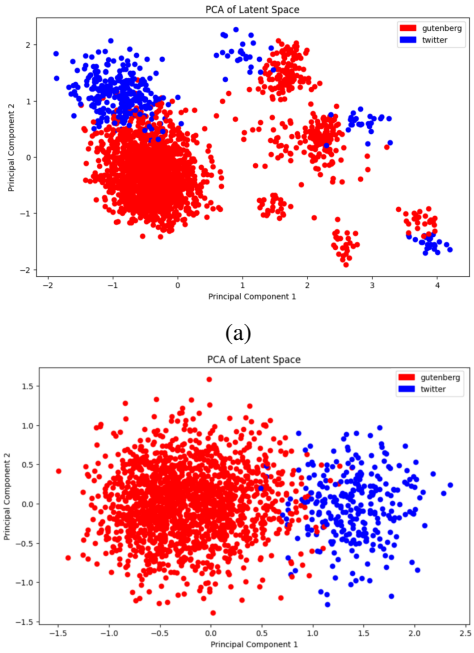
Figure 8: VAE architecture used in this work.



(a)

(b)

Figure 11: Euclidean distance between subreddit weights in the VAE latent space for (a) all data points, and (b) data points belonging to the continuous subset of the latent space.



(a)

(b)

Figure 12: Two-dimensional projection of the VAE latent space trained on the combined Twitter and Gutenberg datasets for (a) the full latent space, and (b) the connected subset of the latent space.

A.5 Interpolation case study

In this section, we present a case study showcasing the functionality of our interpolation process. We pick a data point from the training set in order to compare its interpolated latent representation with its actual latent representation. First, we show a sample result where our approach successfully interpolates latent models. Second, we study how the choice of source models and data samples affect the results. Finally, we show how accumulative interpolation solves some of the issues.

A.5.1 Success Sample

We picked a datapoint from the subreddit r/wallstreetbets. We randomly select source models belonging to other subreddits. Here we pick 3 base models. Figure 13 shows the cross-entropy

results produced by models interpolated using our approach compared to finetuning. As can be seen, our approach is able to interpolate models that outperform finetuning on the given samples. Given the actual position of the datapoint in latent space, we can inspect the operation of our approach, as shown in Figure 14. We also validate these results in Figure 15 by plotting the Euclidean distance in latent space between the interpolated models and the actual location of the target model. These figures show that our approach indeed captures the style of the target author as the number of available text samples increases.

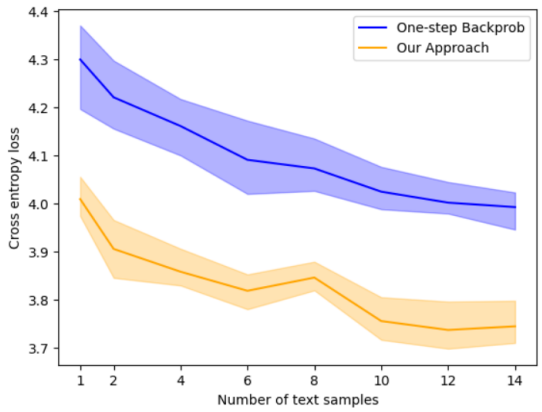


Figure 13: Cross-Entropy Loss Results for the case study data point.

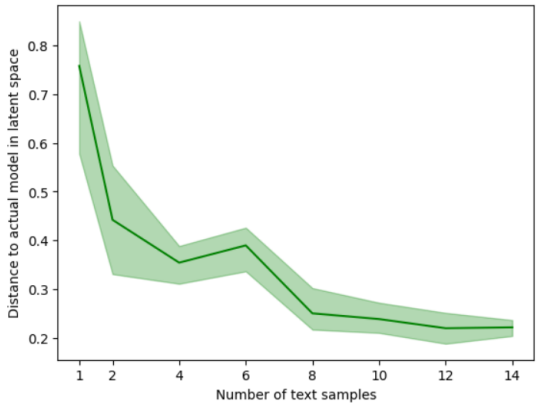


Figure 15: Euclidean distance between interpolated models and the actual model in latent space.

A.5.2 Effect of Text Samples

While a certain author's text corpus can be viewed as a general representation of their style overall, each sample within the corpus can differ in how much it represents that style. Thus, in few-shot settings, there is a risk that a few text samples are not clearly representative of the author and may misguide the system. Figure 16 shows the cross-entropy results for our approach in the same setting but with a different subset of text samples. We can see that the text samples in the range from 6 to 10 push the system away from the target model. This is supported by observing the latent space in Figure 17. Hence, it is crucial that text samples in few-shot settings are chosen to be clean and representative of the target style. We address this issue by performing accumulative interpolation, as discussed in the section A.5.4.

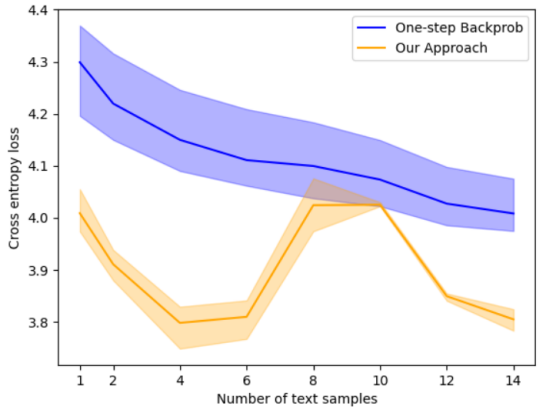


Figure 16: Cross-Entropy Loss Results for different text samples of the case study data point.

A.5.3 Effect of Source Models

Since our approach relies on changes in the source models during finetuning to interpolate finetuned models in the latent space, the choice of source models can have a large impact. In Figure 18, we show the cross-entropy results for source models that belong to the same subreddit as the target model (r/wallstreetbets). Two observations can be made here: (1) the source models perform much better than source models that did not belong to the same subreddit, and (2) the performance of the interpolated models is not as stable as before. Observing the latent space in Figure 19, we find that due to the proximity of the base models to the target model in the latent space, interpolation becomes more difficult due to the instability of the direction of changes in the latent space. We discuss one possible solution for this issue in the next section

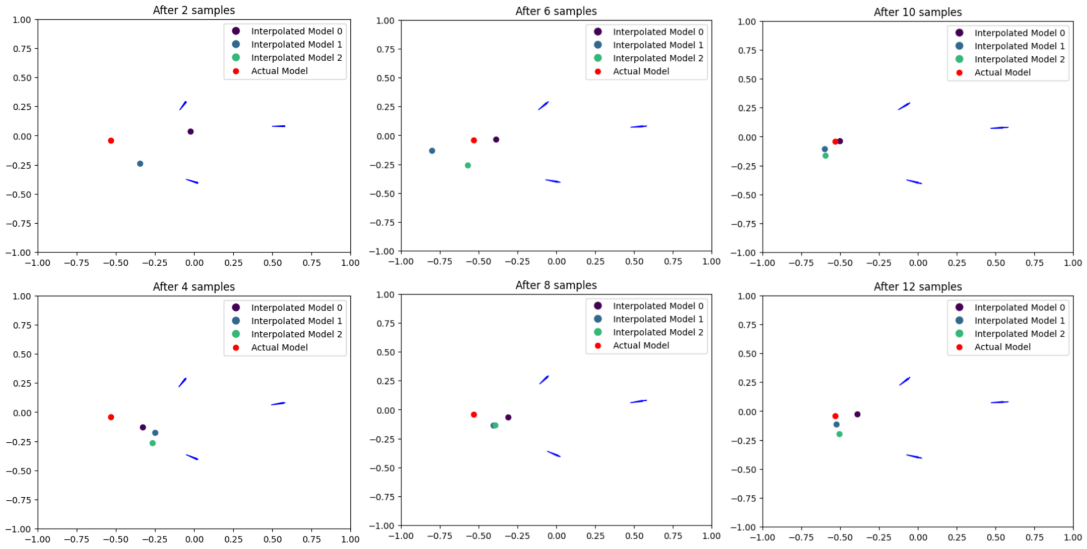


Figure 14: Snapshots of the latent space during interpolation. The red dot represents the location of the target model. The interpolated models successively approach this location as the number of samples increase.

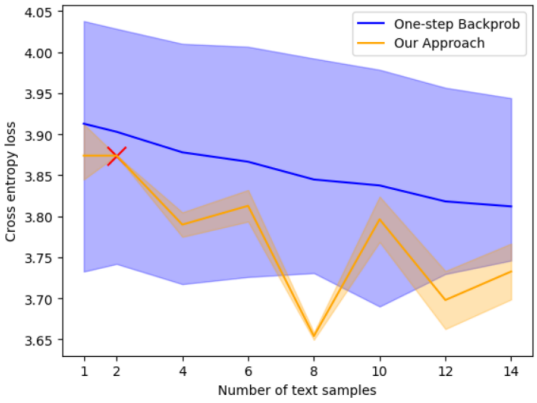


Figure 18: Cross-Entropy Results for the case study data point with source models that belong to the same subreddit. The red X represents failure in interpolation.

A.5.4 Effect of Accumulative Interpolation

Through observing the latent space, we found that poor choice of text samples or source models can lead to unstable results. To address this, we proposed accumulative interpolation as described in section 3.2. Figure 20 shows its effectiveness in mitigating stability issues for various cases.

A.6 Simple interpolation using Reddit

We previously only showed results of applying accumulative interpolation using Reddit. Figure 21 shows the cross-entropy results using simple interpolation. Compared to Figure 2 in section 5.1, simple interpolation leads to more visible variations in the performance of the interpolated models.

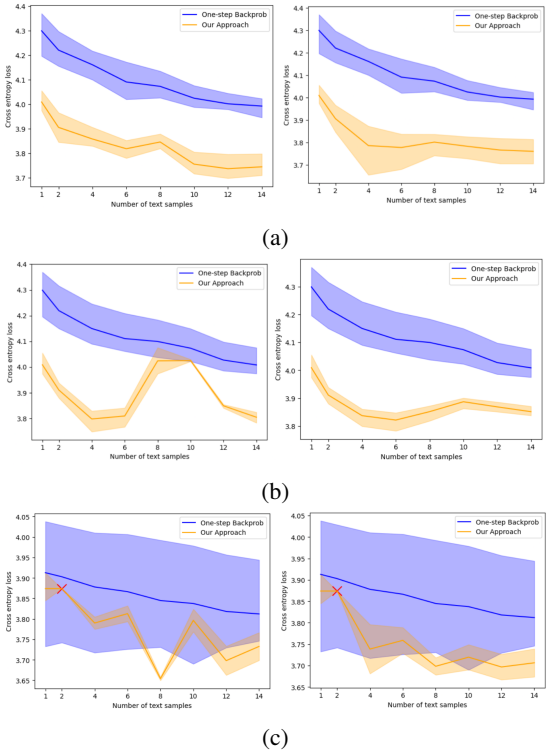


Figure 20: Cross-entropy results using simple interpolation (left) and accumulative interpolation (right) for (a) a good choice of source models and text samples, (b) the same source models but with different text samples, and (c) source models that belong to the same subreddit.

A.7 Reddit results with base models from same subreddit

Figure 22 shows the cross-entropy results of the Reddit dataset when base models are selected to be belonging to the same subreddit as the target

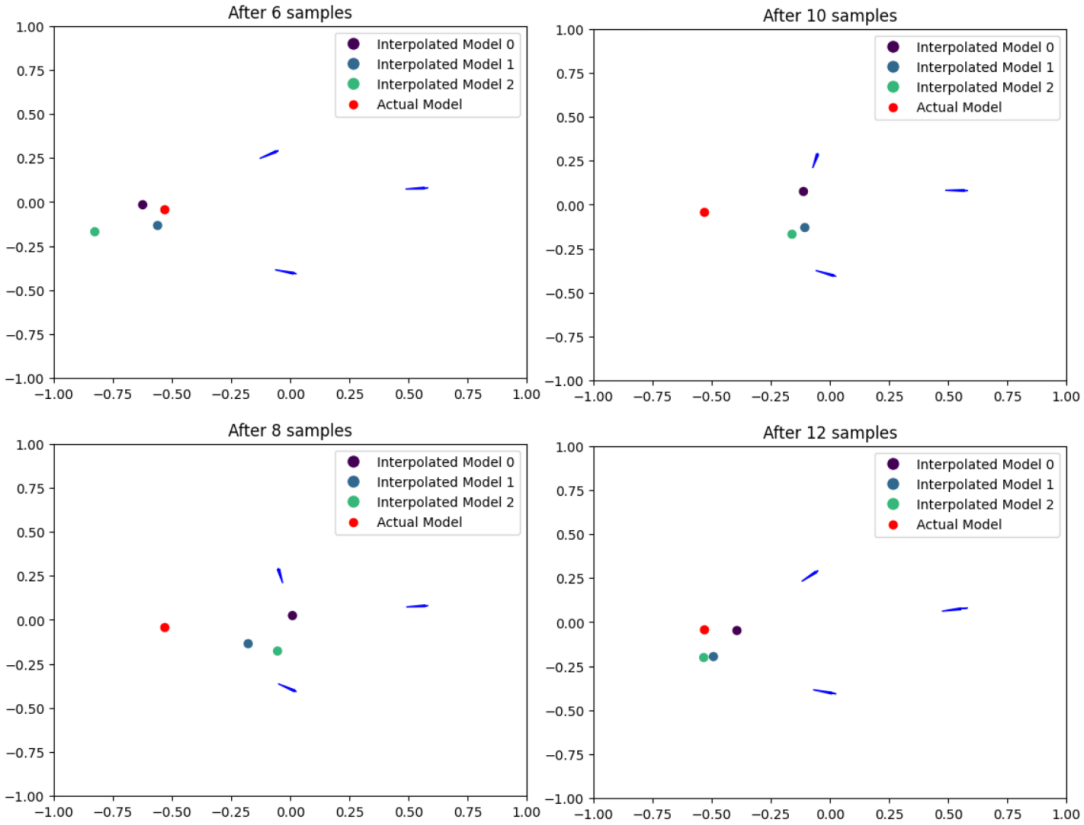


Figure 17: A sequence of snapshots from the latent space during interpolation with a bad choice of text samples.

model. To compare with Figure 2, we show the results for accumulative interpolation. We find that when the base models already belong to the same distribution as the target model, the performance of our approach is almost the same as finetuning.

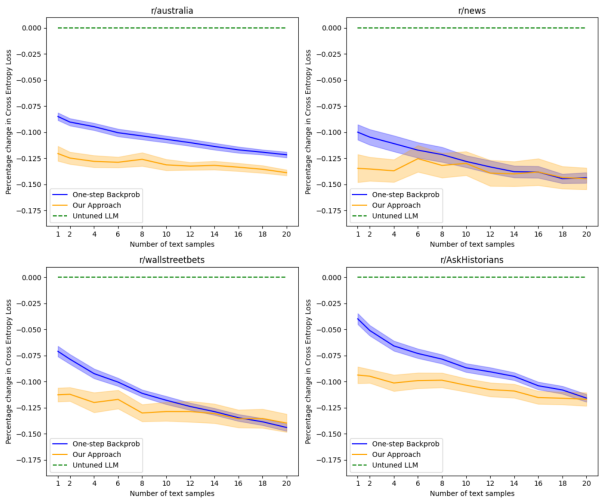


Figure 21: Cross-Entropy Loss Results for Reddit Dataset with Simple Interpolation.

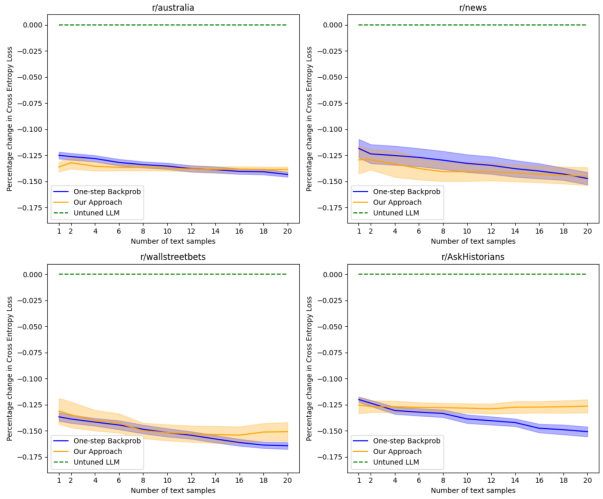


Figure 22: Cross-Entropy Loss for Reddit data with base models that belong to the same subreddit.

A.8 UAR and CISR Results for Twitter and Gutenberg

We report UAR and CISR results for the Twitter and Gutenberg data in Figures 23 and 24. The poorer results compared to Reddit are likely due to both UAR and CISR being trained on Reddit data that is significantly out-of-distribution for the Gutenberg data. This hypothesis is supported by

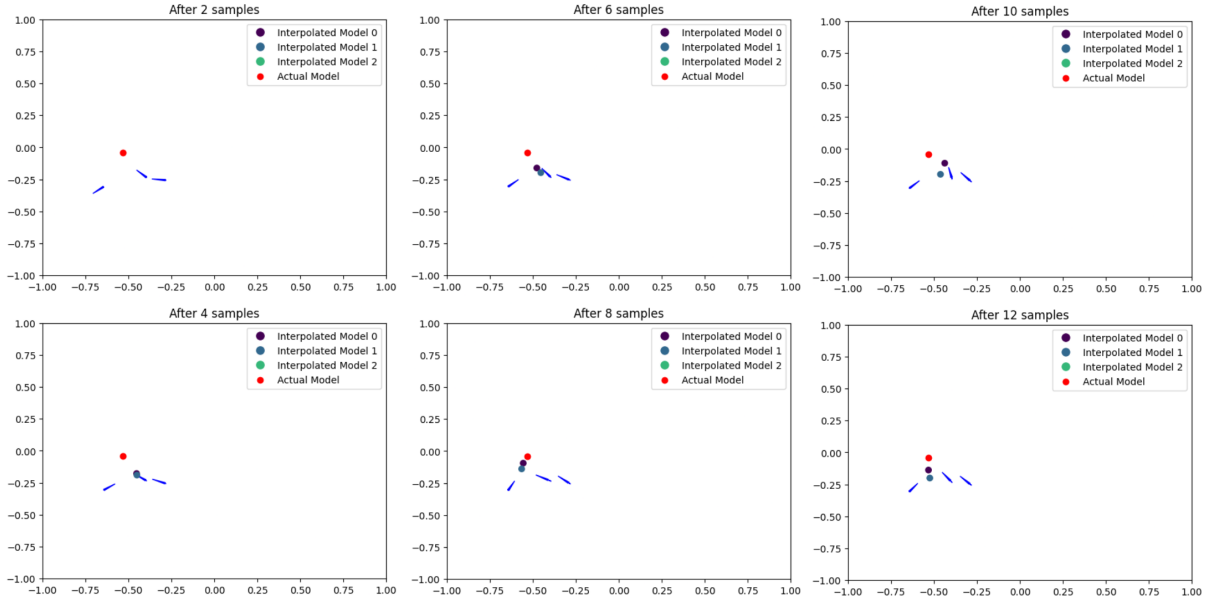


Figure 19: A sequence of snapshots from the latent space during interpolation with source models that belong to the same subreddit as the target model.

the relative consistency of the UAR scores for Twitter which is closer to the style of Reddit.

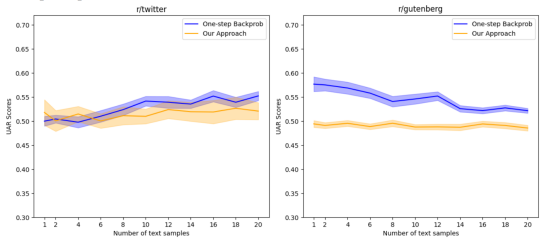


Figure 23: UAR scores for the Twitter and Gutenberg dataset comparing our approach with finetuning.

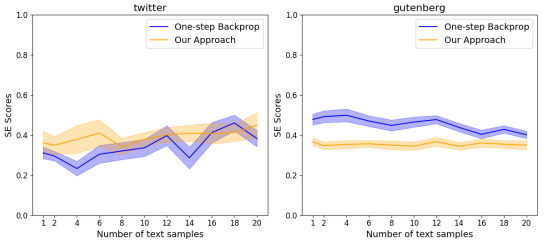


Figure 24: CISR scores for the Twitter and Gutenberg dataset comparing our approach with finetuning.

A.9 GPT-2 Experiment

To test the generality of our approach, we applied our approach using GPT-2 (Radford et al.) as the pretrained model instead of Llama-2. Figure 25 shows the latent space of a VAE trained on the weight deltas of finetuned GPT-2 models using the combined Twitter and Gutenberg dataset. As with Llama-2 (Figure 12b), the VAE is able to differentiate between Twitter and Gutenberg using the weight deltas. Unlike Llama-2, for GPT-2, the resulting latent space was not disconnected and thus we did not have to apply filtering to obtain a connected latent space. This may be due to us being able to finetune GPT-2 directly (i.e. without having to apply LoRA and PCA) due to its smaller size. We hope to study this further in future work.

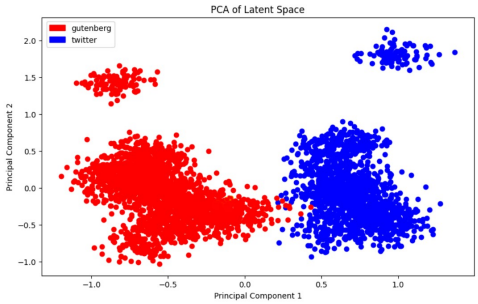


Figure 25: 2D projection of the VAE latent space on the combined Twitter and Gutenberg datasets using GPT-2 as the base pretrained model for finetuning.

A.10 Sample Generations

Finally, we present sample generations for both the Reddit and the Twitter-Gutenberg data in Tables 3 and 4 respectively. Each row shows a randomly chosen sample from the training set and randomly chosen samples from our approach with 2 and 16 samples. Samples in a given row correspond to the same corpus/user.

Training Sample	2 samples	16 samples
r\AskHistory		
You're exactly correct. The army only had significant issues with recruitment during times of natural disasters (like the Antonine Plauge or the Plauge of Justinian) or following military emergencies (like Adrianople), as the immediate needs for large volumes of manpower strained demographic resources.	The first time I read the book, I didn't get it at all. hopefully you will like it better the second time. I have a feeling that the sequel will be just as good as the original. I've read the first book twice, and have the sequ	By the time I was old enough to read The Hobbit, my father had been reading the book to me for years. He knew a bit more about Tolkien's life than I do, but I have always wondered whether the fact that he was born in 189
That is a really good point on using this as Hamlet's characterization. But back on the history topic, what's this "Grand Tour of Europe" of Renaissance Crown Princes?	This week is the 10th anniversary of the 2007–2008 financial crisis. This crisis started in the United States, but quickly spread throughout the world. It was not the first financial crisis that occurred in the world, but it is the most famous	You can find a lot of information on this subject on this website, as well as many other sources, including other websites. Here is an example of a website that provides information on the history of this subject: `http://www.history.org/history/teach/themes/amer`
r\Australia		
Basically you're ok with corporate interests deciding who can eat, and do all kinds of things. Hmm. Do you know what that kind of system of government might be called?	The most important thing to keep in mind when looking at the polls is that a lot of these polls are self-reporting. I think a lot of people are over-reacting to this. They're not going to be able to stop this, it's going to	Well, it's not a big deal. It was just a joke. The only reason I made that comment in the first place was to try and make a point about how people are so quick to point fingers. I'm not actually offended. Just disappointed in the people
You have no idea how popular Home and Away is in Ireland and the UK. Literally everyone watches it, not just teenage girls.	They have no idea. they are a joke and have no idea how to manage the situation. If you were a citizen of the country, they wouldn't be "their" people. You're an a———-, and you don't get to decide whether	The only way to stop this is to stop it at the border. Even then, I don't want to go that far. If you are going to let people in, make them assimilate. They need to learn English and they need to accept our laws. It shouldn't
r\News		
You have a very narrow worldview if you believe everyone that disagrees with you is an idiot.	I think it's just that he has no intention of actually being an MP and he's just going to be another of the "me too" politicians who are going to be around for a decade or two after they've done absolutely nothing.	The first thing I notice is that the new one is a little more aggressive in the front. It looks more like a real car than the current one. I was actually thinking this looks kinda like the current model. The front end looks a little more aggressive, but

That is also abhorrent. Seriously dude, I would never support the torture of anyone, for any reason:	I've done both. I think there are some great places for both. It just depends on what you are looking for in your next step. I love living in a big city. I would have never left NY if I had stayed on my original plan to go to law school and become	It is a shame that you are being ridiculed for your belief. However I would encourage you to think about the possibility that your belief may be wrong. It's okay to be wrong. You are allowed to believe in something even though it is wrong, but you have to
r\wallstreetbets		
You really dont know what youre talking about. Fundamental is anything looking at the intrinsic value, technical does not. Its a boolean, either you are or you arent, there is no other option.	I'm glad that I'm not the only one who gets a bit frustrated with the "give me a number" question. geprüft.com is the best tool for this I've found so far. It's not the only tool, but I'd	The problem is that if you don't get that 30% in a single year, you're probably in the wrong business. And 30% is a big if. It's not impossible. I don't think it's unrealistic, but I
A month from now, they'll have the emergency meeting to put it another month out because f— everyone else	I'm not sure I understand the question, but you should be able to find the answer in the documentation of the package. everybody here has a different definition of "reliable". Some people have very low tolerance for risk and would not be satisfied with a 1% drop over	The only reason I have any faith in it is because I think the market is going to crash, and that it will be a great time to buy. At least I hope so. I'm not sure I'm smart enough to trade options.

Table 3: Sample generations from models output by our method using the Reddit dataset. Each row corresponds to a different user and depicts: an original text sample from that user and random samples generated by a model output from our approach based on 2 and 16 samples from that user respectively.

Training sample	2 Samples	16 Samples
Anna Karenina		
relations with him, which she had hitherto avoided thinking about.	that the man had a large hand and long, thin fingers. To this I could give no other reason than the mere fact of my having seen	It was at the door of the little hut that the boy's face first became intelligible to me. I could not see the eyes
Emma		
not been thinking of him, she had forgotten to think of him—but she was very glad to see him. Mrs. Weston would be at ease. The black mare was blameless;	a time he would have been glad of the company of his wife. But now, he felt, he was too much of a man to be	that I could find no other excuse for my conduct than that I was not a man of principle, but a man of the world. I
Middlemarch		
I have in my hand is an ingenious contrivance—a sort of practical rebus, I may call it: here, you see, it looks like an elegant heart-shaped box, portable—for the pocket;	the door to me, when I was not in the way. They would even take me out to walk, and let me see the world.	I am sure, if the truth were known, that the good man had an interest in my success; but I had never seen him so deeply
Peter Pan		
After all, he was the only man that the Sea-Cook had feared. But Peter had no sinking, he had one feeling only, gladness; and he gnashed his pretty teeth with	his way to a little house, at the end of a small street. nobody was there to greet him, nor any one else. but	the end of the week, and he was still a prisoner. His wife, who had a little money, had come to the end of her
Sense and Sensibility		
her less wretched.	that had fallen upon his shoulder. He lifted it off, and the light of day streamed in upon the room. There was a little	his hand, as he spoke, and a great deal more that was still less complimentary. He then said, "If it is true
Wuthering Heights		
insulted so shockingly by a stranger! She did not comprehend it; and hard work I had to obtain a promise that she would not lay the grievance before her father.	the most beautiful girl he had ever seen. He was not in love with her, but he had fallen a victim to that species of love	and said that he had been a good deal with the family, and was a particular friend of the lady's father, and had been much
Twitter User		
How are you this mornoon?	hopefully they will come out soon. This is really a fantastic and valuable post. Thank you	the most interesting and instructive book I ever read.
Twitter User		
Just havin a chilled Sunday then? I love Sundays for that reason..you can just sit about an be a lazy b and no one cares.	I am a good fellow, a good fellow, I know, I know. I have not a bad heart. I am not bad, no	his mouth. obviously he was a man of no great culture. But I think he must have had a good deal of practical wisdom, and

Table 4: Sample generations from models output from our approach using the Twitter and Gutenberg dataset.