

# STYLEDISTANCE: Stronger Content-Independent Style Embeddings with Synthetic Parallel Examples

Ajay Patel<sup>† \*</sup> Jiacheng Zhu<sup>† \*</sup> Justin Qiu<sup>†</sup> Zachary Horvitz<sup>‡</sup>  
Marianna Apidianaki<sup>†</sup> Kathleen McKeown<sup>‡</sup> Chris Callison-Burch<sup>†</sup>  
University of Pennsylvania<sup>†</sup> Columbia University<sup>‡</sup>

## Abstract

Style representations aim to embed texts with similar writing styles closely and texts with different styles far apart, regardless of content. However, the contrastive triplets often used for training these representations may vary in both style and content, leading to potential content leakage in the representations. We introduce STYLEDISTANCE, a novel approach to training stronger content-independent style embeddings. We use a large language model to create a synthetic dataset of near-exact paraphrases with controlled style variations, and produce positive and negative examples across 40 distinct style features for precise contrastive learning. We assess the quality of our synthetic data and embeddings through human and automatic evaluations. STYLEDISTANCE enhances the content-independence of style embeddings, which generalize to real-world benchmarks and outperform leading style representations in downstream applications. Our model can be found at <https://huggingface.co/StyleDistance/styledistance>.

## 1 Introduction

The most common objective when training text embeddings is to place texts with similar semantics close together in the embedding space (Reimers and Gurevych, 2019). Style representations, by contrast, aim to embed texts with similar writing styles near each other and texts with different styles far apart, regardless of their semantic content (Wegmann et al., 2022). Embeddings are usually trained via contrastive learning, with triplets consisting of an anchor text, a positive text (which should be embedded closely to the anchor), and a negative text (which should be embedded far from the anchor) (Goldberger et al., 2004; Khosla et al., 2020; Schroff et al., 2015). Existing approaches often use

<sup>\*</sup>Denotes equal contribution; direct correspondence to: [ajayp@upenn.edu](mailto:ajayp@upenn.edu)

**Style Feature:** Usage of Active Voice

**Anchor:**

● "I adored the Dolce & Gabbana mauve leather creation with a signature lock."

**Positive & Negative Examples:**

● "I observed the impact rising temperatures had on tropical diseases."

● "The Dolce & Gabbana mauve leather creation with a signature lock was adored by me."

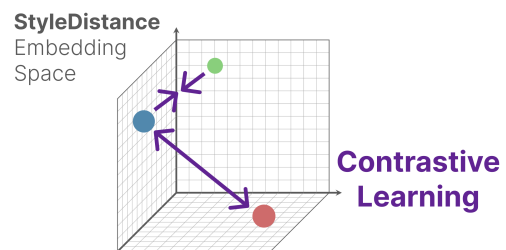


Figure 1: STYLEDISTANCE embeddings are trained using contrastive learning from synthetic parallel (positive and negative) examples representing 40 style features. The illustrated example is for the “Usage of Active Voice” feature.

social media datasets with the assumption that all writing by the same author shares a similar style and that texts by different authors exhibit dissimilar styles (Wegmann et al., 2022). These methods also attempt to minimize content representation in the resulting embeddings. They select a text from the same author on a *different topic* as the positive example, and a text from a different author on the

*same topic* as the negative example, approximating topic similarity using subreddit or conversation metadata. However, these methods are limited by the imperfect nature of data acquired under such assumptions. For example, the same author may write about the same topic even in different subreddits. As a result, these imperfect contrastive triplets do not explicitly control for content, leading to style embeddings with weak content-independence. The “content leakage” caused by such proxy objectives (illustrated in Figure 2) can undermine the effectiveness of style representations in tasks that require strict separation between style and content, such as stylistic analysis, authorship tasks, style transfer steering, and automatic style transfer evaluation. To overcome this, a more controlled approach to style contrastive learning is necessary.

In this paper, we introduce `STYLEDISTANCE`, a novel method for training stronger, content-independent style embeddings which leverages synthetic parallel text examples generated by a large language model (LLM) (OpenAI et al., 2024). By creating near-exact paraphrases with controlled stylistic variations, we produce positive and negative examples across 40 distinct style features. This synthetic dataset, which we call `SYNTHSTEL`, enables more precise contrastive learning (visualized in Figure 1) and is more robust to the content leakage inherent in existing datasets. We evaluate our method on both human and automated benchmarks, measuring the content-independence, quality, and utility of `STYLEDISTANCE` embeddings.

In summary, our primary contributions are:

1. We generate and release `SYNTHSTEL`, a dataset of near-exact paraphrases across 40 distinct style features.
2. We introduce `STYLEDISTANCE`, a new approach to style representation learning which uses synthetic parallel examples with controlled stylistic variations. We release the embedding model as a resource.
3. We demonstrate that `STYLEDISTANCE` significantly improves the content-independence of style embeddings, generalizing effectively to real-world benchmarks of style representation quality and outperforming existing style representations in downstream applications.

## 2 Related Work

**Style Representations** In previous work, style representations were learned from unlabeled texts (Zhu et al., 2022; Dai et al., 2019; Riley et al., 2021, *inter alia*). Due to the lack of parallel datasets covering diverse style features, Hay et al. (2020) and Wegmann et al. (2022) have both recently used authorship as a proxy for style as we discussed previously. Rivera-Soto et al. (2021) trained embeddings to uniquely represent different authors. Although these embeddings capture features representative of authors’ style, they also capture content features due to the lack of control for content-related aspects. Patel et al. (2023) trained LISA, a style vector created by annotating texts with their stylistic features using LLMs. LISA, however, trades off performance to create a vector with interpretable dimensions. The highest quality style representations to date come from approaches that employ a contrastive learning objective (Wegmann et al., 2022). In this paper, we propose leveraging the strengths of generative LLMs to build a dataset that will serve to train strong content-independent style embeddings using a contrastive learning objective. From a practical perspective, style representations are useful in downstream applications such as arbitrary text style transfer (Khan et al., 2023; Horvitz et al., 2023, 2024) where they help steer and guide transfer, and content-independence is important for reducing hallucinations in generations.

**LLMs and Text Style** LLMs are commonly used for style transfer (Reif et al., 2022; Suzgun et al., 2022; Patel et al., 2022; Fisher et al., 2024) and style analysis (Saakyan and Muresan, 2023; Patel et al., 2023). Their strength with style-related tasks has been demonstrated, yet leveraging this knowledge to build strong content-independent style representations has not yet been explored.

## 3 Data Generation

A core component of our proposed `STYLEDISTANCE` approach is a LLM that generates a synthetic dataset with controlled content and style. The dataset is composed of pairs of sentences which are paraphrases of each other (*i.e.* their content is similar) but differ in style. Each pair includes a positive example that showcases a specific stylistic feature (*e.g.*, usage of active voice or emojis) and a corresponding negative example that lacks this feature. In this section, we outline the selected

	Contrastive example used for Wegmann et al. (2022)		Contrastive example used for STYLEDISTANCE ("Usage of Active Voice" Style Feature)	
<b>Anchor:</b>	Awesome game. Took off from work to play it. Halfway through the week playing nonstop. Such a quality product. The Spider-Man game we've needed for years.		I adored the Dolce & Gabbana mauve leather creation with a signature lock.	
<b>Positive:</b>	Such a great Spider-Man game. It really is the best one I've played. Easy to play crazy fun to master. There is so much you can do with Peter's moves...	Same Style Same Content (leads to content leakage in style representations)	I observed the impact rising temperatures had on tropical diseases.	Same Style Different Content
<b>Negative:</b>	No Android version? I guess I'm not getting it	Different Style Different Content (leads to content leakage in style representations)	The Dolce & Gabbana mauve leather creation with a signature lock was adored by me.	Different Style Same Content

Figure 2: In the training triplets used by Wegmann et al. (2022) (left), an anchor text is paired with a positive instance written by the same author, and a negative instance written by a different author, assuming content can be controlled via subreddit/conversation metadata. However, this assumption can fail, leading to uncontrolled content as illustrated. In our dataset used to train STYLEDISTANCE (right), we control for both style and content.

style features and provide implementation details for our synthetic data generation procedure. Additionally, we present evaluations conducted to assess the quality of the synthetic dataset prior to its use in training our style embeddings.

**Style Feature Selection** There is no predefined set of style features, and what features are considered to describe style vs. content can vary across different studies (Jin et al., 2022). For this work, we select 40 style features across 7 broad categories (visualized in Figure 3) which have been addressed in different works on text style (Tausczik and Pennebaker, 2010; Kang and Hovy, 2019; Wegmann and Nguyen, 2021; Jin et al., 2022; Patel et al., 2023). Specifically, we select features for which it is possible to generate both positive and negative examples (e.g., formal/informal, passive/active voice). Since some features can blur the line between style and content (e.g., usage of sarcasm), it might be difficult to generate perfectly parallel positive and negative pairs, with the same content. For these features, we control the generation as much as possible with the aim to obtain near-exact paraphrases. Furthermore, some style features may be impossible to fully remove from a sentence in order to generate a negative example (e.g., usage of articles). For these, we aim for the positive example to contain the feature with higher frequency than the negative example. For more details on all the selected style features, see Appendix A. While these 40 features may not cover the infinite number of styles that may exist, we believe they can serve to learn more primitive features (e.g., use of con-

tractions, use of long words, use of formal style) which may help generalize to more complex styles involving these features (e.g., "professorial style"). We discuss and test this generalization assumption in Section 5.1.

**Generation** For each of the selected features, we generate 100 pairs of positive and negative examples by prompting GPT-4 (OpenAI et al., 2024) with the DataDreamer library (Patel et al., 2024). Each generated pair contains a sentence where the feature is present (positive example), and a paraphrase where the feature is absent or less present (negative example).

Attribute	Values
<b>Topic</b>	A randomly extracted fine-grained topic from C4.
<b>Sentence Length</b>	['10-15 words', '15-20 words', '20-25 words', '25-30 words']
<b>Point of View</b>	['first-person', 'second-person', 'third-person']
<b>Tense</b>	['past', 'present', 'future']
<b>Type of Sentence</b>	['Declarative', 'Semicolon Structure (compound)', 'Question', 'Exclamation']

Table 1: Attributes sampled for the attributed prompt string generation.

Yu et al. (2023) found that LLMs struggle with diversity when prompted to generate text examples. We use their proposed attributed prompt (AttrPrompt) method to ensure generations are sufficiently diverse and varied across basic attributes,

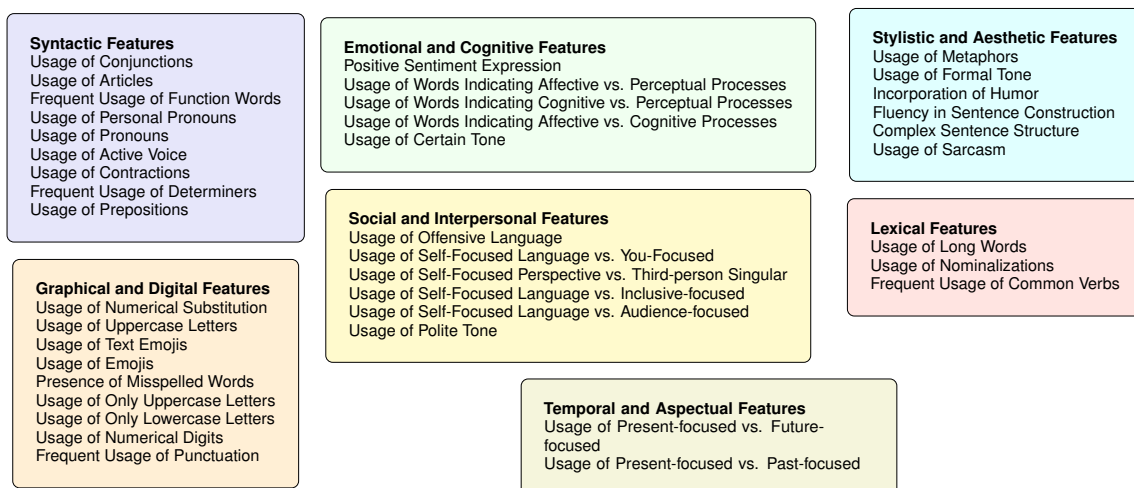


Figure 3: We generate synthetic parallel examples to train STYLEDISTANCE for a wide range of style features in seven linguistic and stylistic categories. Further details on these features can be found in Appendix A.

such as “Sentence Length” and “Type of Sentence”. The method randomly selects values from a defined set of attributes to be included in the prompt, which serve as conditioning for the text generation. In Table 1, we showcase the attributes we sample from in our attributed prompt in order to vary our generations. See Appendix B.2 for details on our full attributed prompt and inference parameters.

For the “Topic” attribute, we sample fine-grained distinct topics for each generation from the C4 corpus (Raffel et al., 2020). We do this by extracting a random sentence from a random document in C4, and we then use a zero-shot prompt (given in Appendix B.1) with GPT-4 to identify the fine-grained topic of that sentence. We employ several heuristics to select sentences from C4 that have desired characteristics: written in English, sufficiently long (greater than 32 words), and consist of natural text rather than formatting text found in some C4 documents. We provide an implementation of these heuristics in our supplementary materials.

We call our final synthetically generated dataset SYNTHSTEL, and create train and test splits using a 90%/10% split stratified by style feature.

**Dataset Evaluation** We conduct a human and an automatic evaluation of the quality of our synthetic dataset for a number of different properties using the test split.

First, we measured the extent to which humans judge our positive examples do contain the desired style feature and our negative examples do not. Note that the appreciation of some style features (such as “Incorporation of Humor”) can be subjective, and some other features (e.g., “Usage of

Articles”) are not fully removed in negative examples but might appear less frequently therein than in positive examples. In spite of these intricacies which made the annotators’ task more difficult, our human evaluation results were strong: 92% of the time, annotators judged the positive and negative labels correct (random chance is 50%). Each instance was annotated by 10 different annotators from a pool of 73 graduate students in a NLP class. We also assessed inter-annotator agreement with Krippendorff’s Alpha (Krippendorff, 2011) and achieved a reliability score of 0.55. For more details about the human annotation, see Appendix C.

We also run automatic evaluations to assess other properties of the dataset:

- **Content Similarity** measures the average semantic similarity between the positive and negative parallel examples (Reimers and Gurevych, 2019).<sup>1</sup>
- **Fluency** measures the average fluency of our examples<sup>2</sup> using a classifier trained on CoLA (Warstadt et al., 2019).
- **Diversity** uses the score proposed by Yang et al. (2024) to measure how different each generated text is from every other in terms of content/topic using semantic similarity.

We compute a baseline for these scores with natural data from the dataset of sentence pairs in Wegmann and Nguyen (2021). The results of these

<sup>1</sup>For evaluations using semantic similarity, we use the all-mpnet-base-v2 model.

<sup>2</sup>We exclude generations for style features that specifically address disfluency.

evaluations can be found in Table 2. Our generated examples fare well in all these aspects: they are topically diverse and fluent, and the similarity inside each pair of positive/negative examples is high. An additional (less direct) evaluation of the quality of our synthetic dataset is proposed in Section 5, where we evaluate the STYLEDISTANCE embeddings that we train on this dataset.

Metric	Baseline	Score
<b>Style Feature Presence</b> (% humans judged correct)	0.50	0.92
<b>Content Similarity</b>	0.88	0.88
<b>Fluency</b>	0.80	0.92
<b>Diversity</b>	0.95	0.91

Table 2: Results of the human and automatic evaluations of our synthetic dataset.

## 4 STYLEDISTANCE

We next describe how we trained STYLEDISTANCE embeddings using our synthetic dataset.

### 4.1 Sampling Contrastive Triplets

After generating 100 pairs of positive and negative examples for each style feature, we construct feature-specific triplets as follows: We select an “anchor” ( $a$ ) and a “positive” example ( $p$ ) from different pairs available for a feature, ensuring that the two examples are **identical in style** but not in content. For the “Usage of Active Voice” example in Figure 1,  $a$  and  $p$  are two active sentences (*I adored ...*, *I observed...*) on different topics. As a negative example ( $n$ ), we use the paraphrase of either  $a$  or  $p$  which does not contain the feature; therefore,  $n$  is always different in style. In the example in Figure 1,  $n$  is the paraphrase of the anchor in passive voice (...*adored by me*).

In terms of **content**,  $n$  is a paraphrase of the anchor  $a$  in half of the triplets, and has different content in the other half. This ensures that the trained model will not only learn to discriminate parallel (paraphrased) texts, but will be able to generalize to texts with different content during inference. This sampling process results in  $\sim 320\text{K}$  unique triplets. The implementation of this simple algorithm is shared in the supplementary materials.

### 4.2 Contrastive Learning Objective

In our final dataset of triplets  $\mathcal{D}$ , each triplet  $(a, p, n) \in \mathcal{D}$  contains an anchor text ( $a$ ), a pos-

itive text ( $p$ ), and a negative text ( $n$ ). We train our embedding model  $f_\theta(\cdot)$  with a triplet loss (with margin  $\alpha$ ) (Schroff et al., 2015):

$$L_t(\theta) = \sum_{(a,p,n) \in \mathcal{D}} [\|f_\theta(a) - f_\theta(p)\|_2^2 - \|f_\theta(a) - f_\theta(n)\|_2^2 + \alpha]_+$$

We use roberta-base as our base model for fine-tuning—the same base model used for the style embeddings in Wegmann et al. (2022)—and perform training with DataDreamer and LoRA (Patel et al., 2024; Hu et al., 2021). We use a margin of 0.1, a learning rate of 1e-4, and a batch size of 512. We further split our training set into a train and validation split (90%/10%) and train using an early stopping patience of 1 epoch. For full details on the training setup, see Appendix D.

### 4.3 Training

We train two versions of our model. STYLEDISTANCE<sub>SYNTH</sub> is fine-tuned only on the synthetic triplets described in Section 4.1. We also train a version using the synthetic triplets for data augmentation (STYLEDISTANCE). In this case, our training set is comprised of 50% natural data—i.e. the triplets used to train the Wegmann et al. (2022) model<sup>3</sup>—and 50% synthetic data. For the augmented model, we hypothesize that mixing in these perfectly parallel synthetic examples will help regularize the model, and discourage it from representing content-related features in favor of style-related ones offering the potential advantages of both approaches: (1) enhanced content-independence, and (2) the ability to capture niche style features in the natural data. We provide a visualization of the learned embedding space of our model after contrastive training using UMAP (McInnes et al., 2018) in Appendix E.

## 5 Evaluation

We propose a direct evaluation of the quality of STYLEDISTANCE embeddings, and an evaluation of their utility in downstream applications. We use other leading style representations like LISA (Patel et al., 2023) and the embeddings from Wegmann et al. (2022) as baselines, and compare them with our style embeddings on the STEL and STEL-or-Content tasks (Wegmann and Nguyen, 2021; Wegmann et al., 2022). We also compare to the LUAR model (Rivera-Soto et al., 2021), an authorship representation model which does not explicitly train content-independent representations but captures

<sup>3</sup>We use the train-conversation split.

Model	Formal		Complex		Numb3r		C'tion		Emoji		Avg	
	STEL	S-o-C	STEL	S-o-C	STEL	S-o-C	STEL	S-o-C	STEL	S-o-C	STEL	S-o-C
<i>Content-Aware Representations</i>												
roberta-base	0.83	0.09	0.73	0.01	0.94	0.13	1.00	0.00	0.98	0.03	<b>0.90</b>	<b>0.05</b>
LUAR	0.80	0.14	0.67	0.00	0.74	0.03	0.77	0.00	0.99	0.02	0.86	0.03
<i>Content-Independent Style Representations</i>												
LISA	0.73	0.05	0.65	0.00	0.85	0.03	0.92	0.00	0.77	0.02	0.71	0.03
Wegmann et al. (2022)	0.83	0.70	0.58	0.27	0.56	0.03	0.96	0.02	0.95	0.07	0.77	0.22
STYLEDISTANCE	0.89	0.72	0.64	0.25	0.84	0.12	1.00	0.20	0.99	0.15	<b>0.87</b>	0.29
STYLEDISTANCE <sub>SYNTH</sub>	0.85	0.73	0.60	0.27	0.71	0.28	0.99	0.22	0.63	0.07	0.76	<b>0.31</b>

Table 3: Accuracy on the STEL/STEL-or-Content (S-o-C) tasks. STYLEDISTANCE leads on both tasks among representations trained for content-independence, and STYLEDISTANCE<sub>SYNTH</sub> generalizes remarkably well to real text data despite being trained only on synthetic data.

Features Tested	Features Used	Formal		Complex		Numb3r		C'tion		Emoji		Avg		Retained Perf.
		STEL	S-o-C	STEL	S-o-C	STEL	S-o-C	STEL	S-o-C	STEL	S-o-C	STEL	S-o-C	
In-Domain	40 out of 40	0.85	0.73	0.60	0.27	0.71	0.28	0.99	0.22	0.63	0.07	0.76	0.31	100%
Out-of-Domain	34 out of 40	0.85	0.67	0.62	0.28	0.56	0.15	0.80	0.00	0.60	0.01	0.68	0.22	65%
Out-of-Distribution	25 out of 40	0.64	0.49	0.65	0.26	0.57	0.11	0.63	0.02	0.68	0.02	0.63	0.18	50%

Table 4: We evaluate how well STYLEDISTANCE<sub>SYNTH</sub> embeddings generalize to unseen style features by ablating features from the synthetic training dataset under three conditions: In-Domain, Out-of-Domain, Out-of-Distribution. We evaluate their performance on the STEL and STEL-or-Content (S-o-C) tasks.

both content and style features in an attempt to represent different authors. LUAR is sometimes used for automatic style transfer evaluation, where we believe a strong style representation would be better suited. We thus choose this task for evaluation.

### 5.1 STEL and STEL-or-Content Evaluation

We first benchmark our embeddings on the STEL and STEL-or-Content tasks that allow for a direct evaluation of the quality of style representations. We briefly describe these tasks below and illustrate examples of these tasks in Appendix F:

- **STEL:** Given two anchor sentences (A1, A2) and two test sentences (S1, S2), STEL measures the ability of an embedding model to pair each test sentence with the anchor sentence that shares the same style based on the cosine similarity of the embeddings of the sentences.
- **STEL-or-Content:** The STEL-or-Content task is similar to STEL but more adversarially challenging, hence better for testing content-independence. In this task, there are again two test sentences (S1, S2) but only a single anchor sentence (A). The test sentence that best matches the *style* of the anchor must be selected; but the incorrect test sentence—which is written in a different style—is a paraphrase

of the anchor with similar *content*. Therefore, in order to succeed on the STEL-or-Content task, a model needs to represent style features stronger than content features.

In their paper, Wegmann and Nguyen (2021) provide a STEL and STEL-or-Content evaluation benchmark over five features with curated natural data. We test our models on this benchmark and present the results in Table 3. Our results are consistent with results reported by Wegmann and Nguyen (2021), who showed that even untrained models like roberta-base can capture style information well, resulting in stronger STEL performance than any of their fine-tuned models. However, the more challenging STEL-or-Content task, which better tests content-independence, shows that only models specifically trained for content-independence are able to capture style features better than content features. Our results indicate that the embeddings generated with our STYLEDISTANCE approach lead over other style representations on both the STEL and STEL-or-Content tasks. Interestingly, we find that STYLEDISTANCE<sub>SYNTH</sub> captures style remarkably well, and manages to generalize to the natural text examples in the evaluation benchmark despite only being trained on our synthetic contrastive triplets. We conclude using synthetic parallel examples during training makes the model more content-independent and helps better capture style.

## 5.2 Generalization Experiment

In an ablation study, we test the ability of our training approach to produce embeddings that can generalize to unseen style features which are not present in the synthetic dataset. We conduct this evaluation by ablating features from the data used to train  $\text{STYLEDISTANCE}_{\text{SYNTH}}$  under three conditions and show the results in Table 4. In the **In-Domain** condition, all 40 style features are included. In the **Out-of-Domain** condition, we exclude synthetic examples corresponding to the five style features in the STEL/STEL-or-Content benchmark.<sup>4</sup> In the **Out-of-Distribution** condition, we further exclude examples for any features similar or indirectly related to the five evaluated features. Details on the exact 15 style features ablated can be found in Appendix H. We compare the Out-of-Domain and Out-of-Distribution performance of  $\text{STYLEDISTANCE}_{\text{SYNTH}}$  on the two tasks to its In-Domain performance, obtained when it was trained on data for all 40 style features. Even in the challenging Out-of-Distribution condition,  $\text{STYLEDISTANCE}_{\text{SYNTH}}$  retains 50% of its performance on the challenging STEL-or-Content task (see the “Retained Perf.” column). This study indicates our training approach generalizes reasonably well to out-of-domain style features which can be composed from style features selected for generation and, to some extent, even to out-of-distribution style features fully outside the selected set.

## 5.3 Synthetic Data for Probing

Our previous experiments demonstrate that training on our synthetic dataset yields strong style embeddings. Next, we investigate whether the synthetic dataset can be used for an entirely different purpose: to probe which specific style features are captured by existing style representations. Our  $\text{SYNTHSTEL}$  dataset allows for the creation of synthetic STEL and STEL-or-Content task instances across a range of 40 style features—much broader than the Wegmann and Nguyen (2021) benchmark where five features were addressed. We use the test split of  $\text{SYNTHSTEL}$  to generate task instances for probing. We examine whether LISA vectors and the Wegmann et al. (2022) style embeddings capture these 40 style features, which  $\text{STYLEDISTANCE}$  models are directly trained to represent. We show average results over all 40 features in Table 5. Per

<sup>4</sup>In our 40 features, there are two separate features for emoji and text emoticons (:-D) so we exclude 6 total features for this condition instead of 5 features, resulting in 34 features.

Model	STEL	S-o-C
LISA	0.79	0.06
Wegmann et al. (2022)	0.76	0.25

Table 5: Results obtained by LISA and the Wegmann et al. (2022) embeddings on STEL and STEL-or-Content instances created from the test split of our  $\text{SYNTHSTEL}$  dataset. See Appendix G for full results.

feature results are provided in Appendix G. Our findings reveal only moderate coverage by LISA and Wegmann et al. (2022), with high variance depending on the evaluated feature (e.g., “Usage of Nominalizations” is poorly captured, with a near-zero STEL-or-Content score for both models). We calculate the mean squared error (MSE) between the STEL and STEL-or-Content scores for the real and synthetic task instances across the five features in the real benchmark, finding an average MSE of 0.039. This small MSE value shows that using synthetic data for probing can reasonably serve to assess which style features are represented by a model without need for manual example curation.

## 5.4 Downstream Evaluation

We next evaluate and/or demonstrate our style embeddings in three downstream applications.

**Authorship Verification** We first test our style embeddings on the authorship verification (AV) task (Koppel and Winter, 2014). Given two documents by unknown authors, the goal of the authorship verification (AV) task is to determine whether they were written by the same author, based on their stylistic similarities and differences (Kocher and Savoy, 2017). We use a series of AV shared task datasets released by PAN in 2011-2015 (Argamon and Juola, 2011; Juola and Stamatatos, 2013; Stamatatos et al., 2014, 2015).<sup>5</sup> Since the two documents may be about different topics, good content-independent style representations would be expected to perform better in the AV task than embeddings that capture content. We calculate the cosine similarity of the two documents using our tested style embedding models with no fine-tuning, to measure their off-the-shelf ability to identify whether two documents were written by the same author and report results using the standard ROC-AUC metric used in AV. In Table 6, we compare

<sup>5</sup>PAN is the “Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection” workshop. No AV shared task was proposed in 2012. (URL: <https://pan.webis.de>)

Model	PAN'11	PAN'13	PAN'14	PAN'15	Avg
LISA	0.55	0.66	0.43	0.64	0.57
Wegmann et al. (2022)	0.65	0.39	0.57	0.63	0.56
STYLEDISTANCE	0.62	0.68	0.58	0.67	<b>0.64</b>

Table 6: ROC-AUC results on the PAN 2011-2015 Authorship Verification (AV) shared tasks.

the performance of STYLEDISTANCE embeddings against LISA and the Wegmann et al. (2022) style embeddings. On average, STYLEDISTANCE outperforms the other representations on AV, demonstrating its effectiveness in representing style.

**Automatic Style Transfer Evaluation** Patel et al. (2022) proposed the LUAR embedding model as an automatic measure for “style transfer accuracy”. This approach was effective and was subsequently adopted for style transfer evaluation (Liu et al., 2024; Horvitz et al., 2023, 2024). However, the LUAR model considers both style and content, hence confounding two aspects of style transfer evaluation—accuracy and meaning preservation—which are typically measured separately. Content-independent style representations would be a better measure for this task. We use the same evaluation dataset of 675 task instances used by Patel et al. (2022) where given an example text of a target author’s style, the task is to discriminate which of two texts (a style transfer output and another actual text by the target author) is written by the target author. We show our results on this task in Table 7. STYLEDISTANCE proves to be a more effective discriminator than all models, including LUAR. All models surpass human performance in distinguishing style transfer outputs. Since automatic style transfer evaluation typically includes a separate score for meaning preservation, using a model like LUAR (which is not content-independent) for measuring style transfer accuracy undermines the rigor of the style transfer accuracy metric. We find that a robust content-independent model like STYLEDISTANCE may enhance automatic style transfer evaluation by: (1) acting as a stronger discriminator, and (2) ensuring style transfer accuracy is assessed independently of meaning preservation.

**Style Transfer Steering** Previous systems, like TinyStyler, have leveraged style embeddings to steer style transfer (Horvitz et al., 2024). While the original TinyStyler system rewrites text by conditioning on Wegmann et al. (2022) embed-

Model	Accuracy
Human	0.37
LUAR	0.38
Wegmann et al. (2022)	0.39
STYLEDISTANCE	<b>0.46</b>

Table 7: Accuracy results on style transfer evaluation.

dings, we demonstrate that STYLEDISTANCE provides an alternative, and reproduce TinyStyler with STYLEDISTANCE embeddings. We showcase an example of an output in Table 8 with more details and results in Appendix I. We will make this version of TinyStyler available as a resource. With this result, we demonstrate STYLEDISTANCE can be used as a simple drop-in replacement for downstream applications in systems where weaker style representations have been previously used.

Source Text (Informal)	<i>"its keeping me up at nite, i have to know what it is"</i>
Wegmann et al. (2022) (→ Formal)	<i>"Have to know what this is, keeping me up at night."</i>
STYLEDISTANCE (→ Formal)	<i>"What is it? It is keeping me up at night."</i>

Table 8: A demonstration of TinyStyler conditioned on STYLEDISTANCE embeddings.

## 6 Conclusion

We introduced STYLEDISTANCE, a novel method for training content-independent style embeddings using synthetic parallel examples. By employing a large language model to generate a dataset of near-exact paraphrases with controlled style variations, we overcome limitations associated with content leakage and imperfect parallel examples in existing style embedding methods. Evaluations using the STEL and STEL-or-Content tasks demonstrate that embeddings trained solely on synthetic examples can capture style extremely well. Additionally, the technique’s ability to generalize to unseen style features indicates its potential to represent a broader range of style attributes beyond those addressed in synthetic data generation. Notably, our results highlight the efficiency of large language models in creating task-specific representations. This approach circumvents the need for manual dataset collection and the reliance on weak implicit assumptions over data, offering a more direct and accurate method for training on the precise task-specific features of interest.



**Model, Data, and Code** We release the STYLEDISTANCE models, the SYNTHSTEL dataset, and our code for other researchers to use at: <https://huggingface.co/StyleDistance/>.

## Limitations

Our approach shows strong results using 40 style features across 7 categories, though it does not fully cover the near-infinite range of possible style variations. The synthetic data may introduce some systematic biases, and we observe occasional repetitive patterns in the generated examples. Nonetheless, our method outperforms existing style representations, and we find that training on our selected 40 features offers strong generalization to unseen styles. Expanding this feature set could further enhance performance. While generating perfectly parallel examples for all style features is challenging—particularly for certain style features that may have overlap with content (Jin et al., 2022)—our model effectively leverages synthetic data to improve content-independence. Additionally, our current focus on sentence-level generations leaves room for future work to explore varying text lengths and multi-sentence style variations, which could further strengthen our approach.

## Ethical Considerations

This work demonstrates the potential of synthetic data for enhancing style embeddings. However, it is important to recognize that the synthetic data generated by large language models may reflect and reinforce existing biases inherent in these models (Patel et al., 2023). While our approach shows significant promise, ongoing efforts should ensure that such synthetic datasets are evaluated for fairness and bias to promote more equitable outcomes.

## Acknowledgements

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

## References

- Shlomo Argamon and Patrick Juola. 2011. PAN11 Author Identification: Attribution. Zenodo. DOI: <https://doi.org/10.5281/zenodo.3713245>.
- Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.
- Ning Dai, Jianze Liang, Xipeng Qiu, and Xuan-Jing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Jillian Fisher, Skyler Hallinan, Ximing Lu, Mitchell Gordon, Zaid Harchaoui, and Yejin Choi. 2024. Styleremix: Interpretable authorship obfuscation via distillation and perturbation of style elements. *Preprint*, arXiv:2408.15666.
- Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. 2004. Neighbourhood components analysis. *Advances in neural information processing systems*, 17.
- Julien Hay, Bich-Liên Doan, Fabrice Popineau, and Ouassim Ait Elhara. 2020. Representation learning of writing style. In *Proceedings of the 6th Workshop on Noisy User-generated Text (W-NUT 2020)*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Zachary Horvitz, Ajay Patel, Chris Callison-Burch, Zhou Yu, and Kathleen McKeown. 2023. Paraguide: Guided diffusion paraphrasers for plug-and-play textual style transfer. *ArXiv*, abs/2308.15459.
- 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. *Preprint*, arXiv:2406.15586.
- 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.
- 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.

- Patrick Juola and Efstathios Stamatatos. 2013. PAN13 Author Identification: Verification. Zenodo. DOI: <https://doi.org/10.5281/zenodo.3715998>.
- Dongyeop Kang and Eduard H. Hovy. 2019. [xs-lue: A benchmark and analysis platform for cross-style language understanding and evaluation](#). *ArXiv*, abs/1911.03663.
- Aleem Khan, Elizabeth Fleming, Noah Schofield, Marcus Bishop, and Nicholas Andrews. 2021. [A deep metric learning approach to account linking](#). *Preprint*, arXiv:2105.07263.
- Aleem Khan, Andrew Wang, Sophia Hager, and Nicholas Andrews. 2023. Learning to generate text in arbitrary writing styles. *arXiv preprint arXiv:2312.17242*.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.
- Mirco Kocher and Jacques Savoy. 2017. A simple and efficient algorithm for authorship verification. *Journal of the Association for Information Science and Technology*, 68(1):259–269.
- Moshe Koppel and Yaron Winter. 2014. Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1):178–187.
- Klaus Krippendorff. 2011. [Computing krippendorff’s alpha-reliability](#).
- Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. [Datasets: A community library for natural language processing](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Shuai Liu, Shantanu Agarwal, and Jonathan May. 2024. [Authorship style transfer with policy optimization](#). *Preprint*, arXiv:2403.08043.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. <https://github.com/huggingface/peft>.
- Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. [Umap: Uniform manifold approximation and projection](#). *Journal of Open Source Software*, 3(29):861.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondrasiuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh,

- Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Ajay Patel, Nicholas Andrews, and Chris Callison-Burch. 2022. Low-resource authorship style transfer with in-context learning. *arXiv preprint arXiv:2212.08986*.
- Ajay Patel, Colin Raffel, and Chris Callison-Burch. 2024. [DataDreamer: A tool for synthetic data generation and reproducible LLM workflows](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3781–3799, Bangkok, Thailand. Association for Computational Linguistics.
- Ajay Patel, Delip Rao, Ansh Kothary, Kathleen MCKeown, and Chris Callison-Burch. 2023. [Learning interpretable style embeddings via prompting LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15270–15290, Singapore. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–6.
- Sudha Rao and Joel Tetreault. 2018. [Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.
- 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.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David C Uthus, and Zarana Parekh. 2021. Textsettr: Few-shot text style extraction and tunable targeted restyling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3786–3800.
- 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.
- Arkadiy Saakyan and Smaranda Muresan. 2023. [Iclef: In-context learning with expert feedback for explainable style transfer](#). *ArXiv*, abs/2309.08583.
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.
- Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A. Sanchez-Perez, and Alberto Barrón-Cedeño. 2014. PAN14 Author Identification: Verification. Zenodo. DOI: <https://doi.org/10.5281/zenodo.3716032>.
- Efstathios Stamatatos, Walter Daelemans, Daelemans and Ben Verhoeven, Patrick Juola, Aurelio López-López, Martin Potthast, and Benno Stein. 2015. PAN15 Author Identification: Verification. Zenodo. DOI: <https://doi.org/10.5281/zenodo.3737563>.
- Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. [Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with](#)

- [small language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.
- Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.
- Anna Wegmann and Dong Nguyen. 2021. Does it capture stel? a modular, similarity-based linguistic style evaluation framework. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7109–7130.
- Anna Wegmann, Marijn Schraagen, Dong Nguyen, et al. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, page 249. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. [Huggingface’s transformers: State-of-the-art natural language processing](#). *CoRR*, abs/1910.03771.
- Yue Yang, Mona Gandhi, Yufei Wang, Yifan Wu, Michael S Yao, Chris Callison-Burch, James C Gee, and Mark Yatskar. 2024. A textbook remedy for domain shifts: Knowledge priors for medical image analysis. *arXiv preprint arXiv:2405.14839*.
- Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. [Large language model as attributed training data generator: A tale of diversity and bias](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 55734–55784. Curran Associates, Inc.
- Kangchen Zhu, Zhiliang Tian, Ruifeng Luo, and Xiaoguang Mao. 2022. [Styleflow: Disentangle latent representations via normalizing flow for unsupervised text style transfer](#). In *International Conference on Language Resources and Evaluation*.

## A Style Features and Definitions

We list all style features selected for our synthetic dataset below along with the positive and negative prompts (used for constructing a full prompt for generating positive and negative examples as shown in Appendix B.2) and definitions (used to help define the style feature to human annotators in the annotation interface in Appendix C).

Style Feature	Positive and Negative Prompts	Style Feature Definition
Usage of Conjunctions	Positive: With conjunctions Negative: Less frequent conjunctions	The "Usage of Conjunctions" text style feature refers to the use of words that connect clauses or sentences. Conjunctions are words like "and", "but", "or", "so", "because", etc. They are used to make sentences longer, more complex, or to show the relationship between different parts of a sentence.
Usage of Numerical Substitution	Positive: With number substitution Negative: Without number substitution	Numerical substitution refers to the practice of replacing certain letters in words with numbers that visually resemble those letters. For example, replacing the letter 'e' with the number '3' in the word 'hello' to make it 'h3llo'. This is a common feature in internet slang and informal digital communication.
Usage of Words Indicating Affective Processes	Positive: Affective processes Negative: Cognitive processes	The text style feature "Usage of Words Indicating Affective Processes" refers to the use of words that express emotions, feelings, or attitudes. These could be words that show happiness, sadness, anger, fear, surprise, or any other emotional state. The presence of such words in a text indicates that the writer is expressing some form of emotional reaction or sentiment.
Usage of Metaphors	Positive: With metaphor Negative: Without metaphor	The "Usage of Metaphors" text style feature refers to the presence of phrases or sentences in the text that describe something by comparing it indirectly to something else. This is often done to make a description more vivid or to explain complex ideas in a more understandable way. For example, saying "time is a thief" is a metaphor because it's not literally true but it helps to convey the idea that time passes quickly and can't be regained.
Usage of Long Words	Positive: Long average word length Negative: Short average word length	The "Usage of Long Words" text style feature refers to the frequency or prevalence of long words, typically those with more than six or seven letters, in a given text. This style feature is often used to measure the complexity or sophistication of the text. If a text has many long words, it is said to have a high usage of long words.
Usage of Uppercase Letters	Positive: With uppercase letters Negative: Without uppercase letters	The usage of uppercase letters as a text style feature refers to the frequency or manner in which capital letters are used in a text. This could be for emphasis, to denote shouting or strong emotions, or to highlight specific words or phrases. It's not just about the start of sentences or proper nouns, but also about other uses of capital letters in the text.
Usage of Articles	Positive: With articles Negative: Less frequent articles	The "Usage of Articles" text style feature refers to how often a text uses words like "a", "an", and "the". These words are called articles and they are used before nouns. This feature measures the frequency of these articles in a given text.
Usage of Text Emojis	Positive: Text Emojis Negative: No Emojis	The text style feature "Usage of Text Emojis" refers to the inclusion of emoticons or smileys in the text. These are combinations of keyboard characters that represent facial expressions or emotions, such as :-D for a big grin or happy face. The presence of these symbols in a text indicates the use of this style feature.
Usage of Nominalizations	Positive: With nominalizations Negative: Without nominalizations	Nominalizations refer to the use of verbs, adjectives, or adverbs as nouns in a sentence. This style feature is often used to make sentences more concise or formal. For example, "the investigation of the crime" is a nominalization of "investigate the crime".
Frequent Usage of Function Words	Positive: With function words Negative: Less frequent function words	The text style feature "Frequent Usage of Function Words" refers to the regular use of words that have little meaning on their own but work in combination with other words to express grammatical relationships. These words include prepositions (like 'in', 'at', 'on'), conjunctions (like 'and', 'but', 'or'), articles (like 'a', 'an', 'the'), and pronouns (like 'he', 'they', 'it').
Usage of Self-Focused Perspective or Words	Positive: Self-focused Negative: Third-person singular	The "Usage of Self-Focused Perspective or Words" text style feature refers to the use of words or phrases that focus on the speaker or writer themselves. This includes the use of first-person pronouns like "I", "me", "my", "mine", and "myself", or statements that express the speaker's personal thoughts, feelings, or experiences.
Usage of Formal Tone	Positive: Formal Negative: Informal	The "Usage of Formal Tone" text style feature refers to the use of language that is polite, impersonal and adheres to established conventions in grammar and syntax. It avoids slang, contractions, colloquialisms, and often uses more complex sentence structures. This style is typically used in professional, academic, or official communications.
Usage of Emojis	Positive: With Emojis Negative: No Emojis	The "Usage of Emojis" text style feature refers to the inclusion of emojis, or digital icons, in a text. Emojis are often used to express emotions, ideas, or objects without using words. If a text contains emojis, it has this style feature.
Usage of Offensive Language	Positive: Offensive Negative: Non-Offensive	The "Usage of Offensive Language" text style feature refers to the presence of words or phrases in the text that are considered rude, disrespectful, or inappropriate. These can include swear words, slurs, or any language that could be seen as insulting or derogatory.
Usage of Present Tense and Present-Focused Words	Positive: Present-focused Negative: Future-focused	The text style feature "Usage of Present Tense and Present-Focused Words" refers to the use of verbs in the present tense and words that focus on the current moment or situation. This means the text is primarily discussing events, actions, or states that are happening now or general truths. It's like the text is talking about what is happening in the present time.

Style Feature Name	Positive and Negative Prompts	Style Feature Definition
Presence of Misspelled Words	Positive: Sentence With a Few Misspelled Words Negative: Normal Sentence	The text style feature "Presence of Misspelled Words" refers to the occurrence of words in a text that are not spelled correctly according to standard dictionary spelling. This could be due to typing errors, lack of knowledge about the correct spelling, or intentional for stylistic or informal communication purposes.
Incorporation of Humor	Positive: With Humor Negative: Without Humor	The "Incorporation of Humor" text style feature refers to the use of language, phrases, or expressions in a text that are intended to make the reader laugh or feel amused. This could include jokes, puns, funny anecdotes, or witty remarks. It's all about adding a touch of comedy or light-heartedness to the text.
Usage of Personal Pronouns	Positive: With personal pronouns Negative: Less frequent pronouns	The "Usage of Personal Pronouns" text style feature refers to the use of words in a text that refer to a specific person or group of people. These words include "I", "you", "he", "she", "it", "we", and "they". The presence of these words in a text can indicate a more personal or direct style of communication.
Fluency in Sentence Construction	Positive: Fluent sentence Negative: Disfluent sentence	"Fluency in Sentence Construction" refers to the smoothness and ease with which sentences are formed and flow together. It involves using correct grammar, appropriate vocabulary, and logical connections between ideas. A text with this feature would read smoothly, without abrupt changes or awkward phrasing.
Usage of Only Uppercase Letters	Positive: All Upper Case Negative: Proper Capitalization	The usage of only uppercase letters style feature refers to the practice of writing all the letters in a text in capital letters. This means that every single letter in the text, whether at the beginning, middle, or end of a sentence, is capitalized. It's like the 'Caps Lock' key on your keyboard is always turned on while typing the text.
Usage of Self-Focused Perspective or Words	Positive: Self-focused Negative: Inclusive-focused	The "Usage of Self-Focused Perspective or Words" text style feature refers to the use of words or phrases that focus on the speaker or writer themselves. This includes the use of first-person pronouns like "I", "me", "my", "mine", and "myself", or statements that express the speaker's personal thoughts, feelings, or experiences.
Usage of Pronouns	Positive: With pronouns Negative: Less frequent pronouns	The "Usage of Pronouns" text style feature refers to the frequency and types of pronouns used in a text. Pronouns are words like 'he', 'she', 'it', 'they', 'we', 'you', 'I', etc., that stand in place of names or nouns in sentences. This feature can indicate the level of personalization, formality, or perspective in a text.
Usage of Words Indicating Cognitive Processes	Positive: Cognitive process Negative: Perceptual process	The text style feature "Usage of Words Indicating Cognitive Processes" refers to the use of words that show thinking or mental processes. These words can express understanding, knowledge, belief or doubt. For example, words like 'think', 'know', 'believe', 'understand' are used to indicate cognitive processes.
Complex Sentence Structure	Positive: Complex Negative: Simple	The "Complex Sentence Structure" text style feature refers to sentences that contain multiple ideas or points, often connected by conjunctions (like 'and', 'but', 'or') or punctuation (like commas, semicolons). These sentences often include dependent clauses, which are parts of the sentence that can't stand alone as a complete thought, alongside independent clauses, which can stand alone. In simpler terms, if a sentence has more than one part and these parts are linked together in a way that they give more detailed information or express multiple thoughts, it has a complex sentence structure.
Positive Sentiment Expression	Positive: Positive Negative: Negative	Positive Sentiment Expression is a text style feature that refers to the use of words, phrases, or expressions that convey a positive or optimistic viewpoint or emotion. This could include expressions of happiness, joy, excitement, love, or any other positive feelings. The text is considered to have this feature if it makes the reader feel good or positive after reading it.
Usage of Numerical Digits	Positive: With digits Negative: Less frequent digits	The "Usage of Numerical Digits" text style feature refers to the presence and use of numbers in a text. This includes any digit from 0-9 used alone or in combination to represent quantities, dates, times, or any other numerical information.
Usage of Words Indicating Affective Process	Positive: Affective process Negative: Perceptual process	The "Usage of Words Indicating Affective Process" text style feature refers to the use of words that express emotions, feelings, or attitudes. These words can show positive or negative sentiments, like happiness, anger, love, or hate. If a text uses a lot of these words, it means the writer is expressing a lot of emotion or personal feelings.
Usage of Active Voice	Positive: Active Negative: Passive	The usage of active voice in a text style feature refers to sentences where the subject performs the action stated by the verb. In other words, the subject is active and directly involved in the action. For example, in the sentence "The cat chased the mouse", 'the cat' is the subject that is actively doing the chasing.
Usage of Only Lowercase Letters	Positive: All Lower Case Negative: Proper Capitalization	The style feature "usage of only lowercase letters" refers to the practice of writing all words in a text with small letters only, without using any capital letters. This means that even the first word of a sentence, proper nouns, or the pronoun 'I' are not capitalized. It's like writing a whole text without ever pressing the shift key on your keyboard.
Frequent Usage of Common Verbs	Positive: With common verbs Negative: Less frequent common verbs	The text style feature "Frequent Usage of Common Verbs" refers to the regular use of basic action words in a text. These are often simple, everyday verbs that are widely used in language, such as 'is', 'have', 'do', 'say', 'go', etc. If a text frequently uses these common verbs, it has this style feature.
Usage of Prepositions	Positive: With prepositions Negative: Less frequent prepositions	The "Usage of Prepositions" text style feature refers to the use of words that link nouns, pronouns, or phrases to other words within a sentence. These words often indicate location, direction, time, or manner. Examples of prepositions include words like "in", "at", "on", "over", "under", "after", and "before".

Style Feature Name	Positive and Negative Prompts	Style Feature Definition
Usage of Self-Focused Language	Positive: Self-focused Negative: Audience-focused	The "Usage of Self-Focused Language" text style feature refers to the use of words or phrases that focus on the speaker or writer themselves. This includes the use of first-person pronouns like "I", "me", "my", "mine", and "myself". It's a way of writing or speaking where the person is often referring to their own thoughts, feelings, or experiences.
Usage of Certain Tone	Positive: Certain Negative: Uncertain	This text style feature refers to the use of a confident tone in writing, where the author avoids using uncertain words or phrases such as 'I think', 'might', or 'seems'. This results in a text that appears more assertive and sure of the information being presented.
Usage of Present-Focused Tense and Words	Positive: Present-focused Negative: Past-focused	The "Usage of Present-Focused Tense and Words" text style feature refers to the use of verbs in the present tense and words that focus on the current moment or situation. This means the text is primarily discussing events, actions, or states that are happening right now or generally true.
Usage of Sarcasm	Positive: With sarcasm Negative: Without sarcasm	The "Usage of Sarcasm" text style feature refers to the presence of statements or expressions in the text that mean the opposite of what they literally say, often used to mock or show irritation. This style is often characterized by irony, ridicule, or mockery, and is used to express contempt or to criticize something or someone in a humorous way.
Usage of Self-Focused Perspective or Words	Positive: Self-focused Negative: You-focused	The "Usage of Self-Focused Perspective or Words" text style feature refers to the use of words or phrases that focus on the speaker or writer themselves. This includes the use of first-person pronouns like "I", "me", "my", "mine", and "myself", or statements that express the speaker's personal thoughts, feelings, or experiences.
Frequent Usage of Punctuation	Positive: With frequent punctuation Negative: Less Frequent punctuation	The text style feature "Frequent Usage of Punctuation" refers to the regular and abundant use of punctuation marks such as commas, periods, exclamation points, question marks, etc., in a piece of text. This style feature is present when the writer often uses these symbols to structure their sentences, express emotions, or emphasize certain points.
Usage of Polite Tone	Positive: Polite Negative: Impolite	The "Usage of Polite Tone" text style feature refers to the use of respectful and considerate language in a text. This can include using words like 'please', 'thank you', or phrases that show deference or respect to the reader. It's about making the text sound courteous and respectful, rather than demanding or rude.
Usage of Contractions	Positive: With contractions Negative: Without contractions	The "Usage of Contractions" text style feature refers to the use of shortened forms of words or phrases in a text. These are typically formed by omitting certain letters or sounds and replacing them with an apostrophe, such as "don't" for "do not" or "I'm" for "I am". If a text frequently uses such shortened forms, it has this style feature.
Frequent Usage of Determiners	Positive: With determiners Negative: Less frequent determiners	The text style feature "Frequent Usage of Determiners" refers to the regular use of words that introduce a noun and give information about its quantity, proximity, definiteness, etc. These words include 'the', 'a', 'an', 'this', 'that', 'these', 'those', 'my', 'your', 'his', 'her', 'its', 'our', 'their'. If a text often uses such words, it has this style feature.

Table 9: The style features selected for synthetic data generation in this work.

## B Generation Prompts and Details

Below we detail the structure of our prompts and inference parameters used for synthetic data generation.

### B.1 Extracting Topics from C4

We select random sentences from C4 (Raffel et al., 2020), and extract the fine-grained topic of the sentence using GPT-4 (OpenAI et al., 2024) and the zero-shot prompt shown below. We perform sampling with temperature = 1.0 and top\_p = 0.0.

```
What is the fine-grained topic of the following text: {sentence} Only return the topic.
```

The fine-grained topic is then used as part of the attributed prompt described in Section B.2 to ensure diversity in the generations.

### B.2 Generating Positive and Negative Example Sentences for Each Style Feature

For each style feature, we generate positive and negative parallel examples using a zero-shot prompt and the attributed prompt (AttrPrompt) method (Yu et al., 2023) with GPT-4 to create diverse and realistic synthetic examples. To generate many examples per style feature, we use a new unused topic extracted from C4 in the prompt each time, we randomly sample a new permutation of the attributes in the prompt, and we perform sampling with temperature = 1.0 and top\_p = 1.0. We demonstrate an example below for the “Usage of Active Voice” style feature.

```
Generate a pair of active and passive sentences with the following attributes:
```

1. Topic: {topic}
2. Length: {sentence\_length}
3. Point of view: {point\_of\_view}
4. Tense: {tense}
5. Type of Sentence: {sentence\_type}

```
Ensure that the generated sentences meet the following conditions:
```

1. There is no extra information in one sentence that is not in the other.
  2. The difference between the two sentences is subtle.
  3. The two sentences have the same length.
- ```
{special_conditions_for_style_feature}
```

```
Use Format:
```

```
Active: [sentence]  
Passive: [sentence]
```

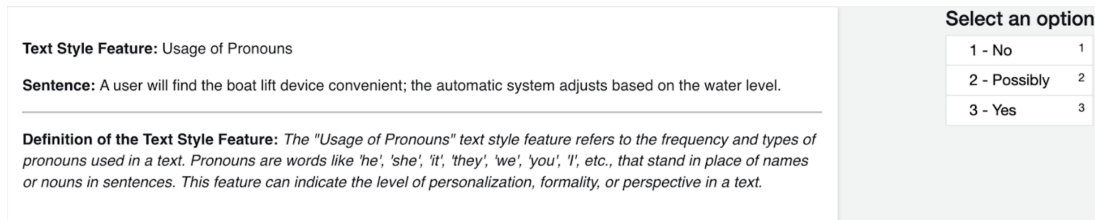
```
Your response should only consist of the two sentences, without quotation marks.
```

For the exact prompts for each style feature, see the code in our supplementary materials for this work.



## C Human Annotation Details

We provide an example of a task instance in our annotation interface. Human annotators were asked to rate whether the style feature was present or not in the sentence, with the option to also select "Possibly" if the annotator was unsure (instructed to use sparingly). We provide annotators with a definition of each style feature as well.



The screenshot shows a user interface for human annotation. On the left, there is a text box containing the following information:

**Text Style Feature:** Usage of Pronouns

**Sentence:** A user will find the boat lift device convenient; the automatic system adjusts based on the water level.

---

**Definition of the Text Style Feature:** *The "Usage of Pronouns" text style feature refers to the frequency and types of pronouns used in a text. Pronouns are words like 'he', 'she', 'it', 'they', 'we', 'you', 'I', etc., that stand in place of names or nouns in sentences. This feature can indicate the level of personalization, formality, or perspective in a text.*

On the right side of the interface, there is a table titled "Select an option" with three rows:

| Select an option |   |
|------------------|---|
| 1 - No           | 1 |
| 2 - Possibly     | 2 |
| 3 - Yes          | 3 |

Figure 4: The annotation interface used for human annotation.

We used a population of graduate students taking a class on natural language processing as the annotators. Each task instance was annotated by 10 distinct human annotators. We assign a score of 0 to “No”, 0.5 to “Possibly”, and 1 to “Yes”. We average the scores from all 10 annotators assigned to each task instance. We consider to have agreement for a positive example if the average score is  $\geq 0.5$ , and for a negative example if the average score is  $< 0.5$ .

We measure inter-annotator agreement using Krippendorff’s Alpha (Krippendorff, 2011) which indicates moderate agreement of 0.55. As a more easily interpretable measure of agreement between annotators, for each task instance, we also find, on average, around 8 out of the 10 annotators annotated in agreement on whether a style feature was present or not in the text.

## D Training Details

| Hyperparameter           | Value                                               |
|--------------------------|-----------------------------------------------------|
| Model                    | Facebook/roberta-base                               |
| Hardware                 | 4x or 8x NVIDIA RTX A6000                           |
| Distributed Protocol     | PyTorch FSDP                                        |
| Data Type                | torch.bfloat16                                      |
| Loss Function            | TripletLoss (Schroff et al., 2015)                  |
| Triplet Loss Margin      | 0.1                                                 |
| LoRA (Hu et al., 2021)   | all-linear, r=8<br>lora_alpha=8<br>lora_dropout=0.0 |
| Optimizer                | adamw_torch                                         |
| Learning Rate            | 1e-4                                                |
| Weight Decay             | 0.01                                                |
| Learning Rate Scheduler  | linear                                              |
| Warmup Steps             | 0                                                   |
| Batch Size               | 512                                                 |
| Train-Validation Split   | 90/10%                                              |
| Early Stopping Threshold | 0.0                                                 |
| Early Stopping Patience  | 1 epoch                                             |

Table 10: Hyperparameters selected for contrastive learning training experiments.

More exact training details can be found in the source code provided in the supplementary materials for this work.

## E Visualization of STYLEDISTANCE Embedding Space

We compare the embedding space of [Wegmann et al. \(2022\)](#) and STYLEDISTANCE on informal/formal texts from GYAFC<sup>6</sup> ([Rao and Tetreault, 2018](#)) in Figure 5 below.

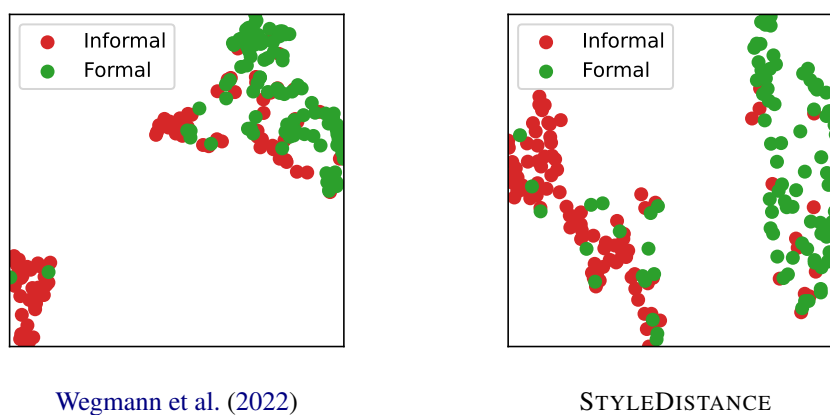


Figure 5: UMAP visualizations of style embeddings from [Wegmann et al. \(2022\)](#) and STYLEDISTANCE on  $n = 100$  random parallel formal/informal examples. [Wegmann et al. \(2022\)](#) forms two distinct clusters of informal texts, making many informal examples distant in the embedding space despite sharing the same style.

---

<sup>6</sup>Grammarly’s Yahoo Answers Formality Corpus which contains 110K informal/formal sentence pairs.

## F STEL and STEL-or-Content Task Visualization

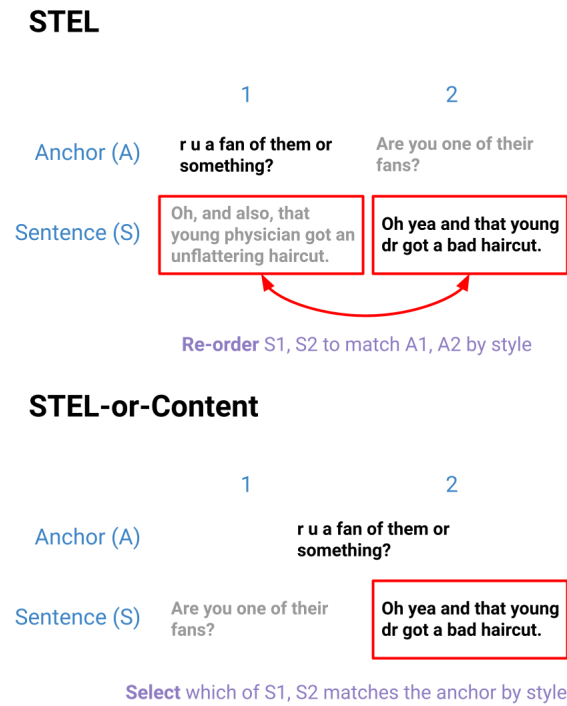


Figure 6: A visualization of the STEL and STEL-or-Content task evaluation we describe in Section 5.1.

## G STEL and STEL-or-Content Results on SYNTHSTEL

| Style Feature                                                | LISA        |             | Wegmann et al. (2022) |             | STYLEDISTANCE |             | STYLEDISTANCE <sub>SYNTH</sub> |             |
|--------------------------------------------------------------|-------------|-------------|-----------------------|-------------|---------------|-------------|--------------------------------|-------------|
|                                                              | STEL        | S-o-C       | STEL                  | S-o-C       | STEL          | S-o-C       | STEL                           | S-o-C       |
| Usage of Polite Tone                                         | 0.93        | 0.18        | 0.76                  | 0.53        | 1.00          | 1.00        | 0.78                           | 1.00        |
| Incorporation of Humor                                       | 0.78        | 0.27        | 0.76                  | 0.31        | 1.00          | 1.00        | 0.82                           | 0.89        |
| Usage of Sarcasm                                             | 0.60        | 0.04        | 0.87                  | 0.11        | 0.98          | 1.00        | 0.53                           | 0.56        |
| Usage of Metaphors                                           | 0.91        | 0.02        | 0.53                  | 0.16        | 0.98          | 0.73        | 0.58                           | 0.82        |
| Usage of Offensive Language                                  | 1.00        | 0.40        | 0.80                  | 0.13        | 1.00          | 1.00        | 0.49                           | 0.93        |
| Positive Sentiment Expression                                | 1.00        | 0.51        | 0.51                  | 0.04        | 0.96          | 1.00        | 0.47                           | 0.53        |
| Usage of Active Voice                                        | 0.64        | 0.00        | 0.64                  | 0.07        | 1.00          | 1.00        | 0.91                           | 1.00        |
| Usage of Certain Tone                                        | 1.00        | 0.20        | 0.60                  | 0.00        | 0.87          | 1.00        | 0.87                           | 0.93        |
| Usage of Self-Focused Language vs. Inclusive-focused         | 0.87        | 0.00        | 0.64                  | 0.00        | 0.76          | 1.00        | 0.69                           | 0.89        |
| Usage of Self-Focused Language vs. You-Focused               | 0.96        | 0.00        | 0.73                  | 0.04        | 0.78          | 1.00        | 0.60                           | 0.80        |
| Usage of Self-Focused Language vs. Audience-focused          | 0.73        | 0.00        | 1.00                  | 0.16        | 0.91          | 1.00        | 0.67                           | 1.00        |
| Usage of Self-Focused Perspective vs. Third-person Singular  | 0.76        | 0.00        | 0.44                  | 0.02        | 0.73          | 1.00        | 0.60                           | 0.82        |
| Usage of Personal Pronouns                                   | 0.64        | 0.00        | 0.56                  | 0.04        | 0.82          | 1.00        | 0.80                           | 0.78        |
| Usage of Present-focused vs. Future-focused                  | 1.00        | 0.00        | 0.56                  | 0.02        | 0.58          | 1.00        | 0.62                           | 0.80        |
| Usage of Present-focused vs. Past-focused                    | 0.89        | 0.00        | 0.60                  | 0.04        | 0.89          | 1.00        | 0.47                           | 0.91        |
| Usage of Words Indicating Affective vs. Cognitive Processes  | 1.00        | 0.07        | 0.87                  | 0.09        | 0.64          | 1.00        | 0.69                           | 1.00        |
| Usage of Words Indicating Affective vs. Perceptual Processes | 0.69        | 0.02        | 0.67                  | 0.00        | 1.00          | 1.00        | 1.00                           | 1.00        |
| Usage of Words Indicating Cognitive vs. Perceptual Processes | 0.76        | 0.04        | 0.56                  | 0.04        | 0.91          | 1.00        | 0.71                           | 0.69        |
| Usage of Articles                                            | 0.69        | 0.00        | 0.69                  | 0.02        | 0.89          | 1.00        | 0.82                           | 0.93        |
| Fluency in Sentence Construction                             | 0.60        | 0.00        | 0.91                  | 0.62        | 0.98          | 1.00        | 1.00                           | 1.00        |
| Frequent Usage of Function Words                             | 0.58        | 0.02        | 0.56                  | 0.24        | 0.82          | 1.00        | 0.80                           | 0.96        |
| Frequent Usage of Common Verbs                               | 0.67        | 0.00        | 0.82                  | 0.07        | 0.78          | 0.89        | 0.49                           | 0.89        |
| Usage of Pronouns                                            | 0.96        | 0.00        | 0.53                  | 0.24        | 0.78          | 1.00        | 0.62                           | 0.80        |
| Usage of Prepositions                                        | 0.67        | 0.00        | 0.78                  | 0.13        | 0.73          | 1.00        | 0.82                           | 0.71        |
| Frequent Usage of Determiners                                | 0.53        | 0.00        | 0.69                  | 0.07        | 0.93          | 1.00        | 0.53                           | 0.98        |
| Usage of Conjunctions                                        | 0.49        | 0.00        | 0.64                  | 0.31        | 0.67          | 1.00        | 0.67                           | 0.96        |
| Usage of Nominalizations                                     | 0.58        | 0.00        | 0.96                  | 0.07        | 0.80          | 1.00        | 0.56                           | 1.00        |
| Usage of Long Words                                          | 1.00        | 0.00        | 0.91                  | 0.44        | 1.00          | 1.00        | 0.64                           | 0.91        |
| Usage of Numerical Digits                                    | 0.58        | 0.00        | 0.58                  | 0.02        | 0.93          | 0.80        | 0.58                           | 0.62        |
| Usage of Uppercase Letters                                   | 0.78        | 0.00        | 1.00                  | 0.98        | 1.00          | 1.00        | 1.00                           | 1.00        |
| Frequent Usage of Punctuation                                | 0.56        | 0.00        | 0.93                  | 0.58        | 0.78          | 1.00        | 0.67                           | 0.80        |
| Usage of Formal Tone                                         | 0.89        | 0.00        | 0.98                  | 0.36        | 1.00          | 1.00        | 0.64                           | 0.98        |
| Complex Sentence Structure                                   | 0.56        | 0.00        | 0.60                  | 0.16        | 0.51          | 0.76        | 0.73                           | 0.87        |
| Usage of Contractions                                        | 0.71        | 0.02        | 1.00                  | 0.31        | 1.00          | 0.89        | 1.00                           | 0.93        |
| Usage of Numerical Substitution                              | 0.81        | 0.00        | 0.90                  | 0.04        | 0.90          | 0.87        | 0.90                           | 1.00        |
| Usage of Only Lowercase Letters                              | 0.89        | 0.00        | 1.00                  | 1.00        | 1.00          | 1.00        | 1.00                           | 1.00        |
| Usage of Only Uppercase Letters                              | 0.98        | 0.33        | 0.87                  | 0.18        | 1.00          | 1.00        | 1.00                           | 1.00        |
| Presence of Misspelled Words                                 | 1.00        | 0.24        | 1.00                  | 0.91        | 1.00          | 1.00        | 1.00                           | 0.96        |
| Usage of Text Emojis                                         | 1.00        | 0.02        | 0.91                  | 0.78        | 1.00          | 1.00        | 1.00                           | 1.00        |
| Usage of Emojis                                              | 1.00        | 0.00        | 1.00                  | 0.47        | 1.00          | 1.00        | 1.00                           | 1.00        |
| <b>Average</b>                                               | <b>0.79</b> | <b>0.06</b> | <b>0.76</b>           | <b>0.25</b> | <b>0.88</b>   | <b>0.97</b> | <b>0.74</b>                    | <b>0.89</b> |

Table 11: STEL and STEL-or-Content results for top-performing models on the SYNTHSTEL test split. This table shows the performance variations and coverage of LISA and Wegmann et al. (2022) embeddings for the 40 different style features found in the dataset. After training STYLEDISTANCE models on the SYNTHSTEL train split, we observe strong coverage of these 40 style features as expected, demonstrating the successful distillation of the LLM’s strong style knowledge into a more efficient representation model (Hinton et al., 2015).

## **H Style Feature Ablation Details**

For the generalization experiment and ablation results we demonstrate in Table 4, we list the style features ablated (removed from the training data) for the Out-of-Domain and Out-of-Distribution conditions.

### **Out-of-Domain:**

1. Usage of Formal Tone
2. Usage of Contractions
3. Usage of Numerical Substitution
4. Complex Sentence Structure
5. Usage of Text Emojis
6. Usage of Emojis

### **Out-of-Distribution:**

1. Usage of Formal Tone
2. Usage of Polite Tone
3. Fluency in Sentence Construction
4. Usage of Only Uppercase Letters
5. Usage of Only Lowercase Letters
6. Incorporation of Humor
7. Usage of Sarcasm
8. Usage of Contractions
9. Usage of Numerical Substitution
10. Usage of Numerical Digits
11. Complex Sentence Structure
12. Usage of Long Words
13. Usage of Text Emojis
14. Usage of Emojis
15. Presence of Misspelled Words

## I Style Transfer Performance

TinyStyler is a style transfer system that reconstructs texts by conditioning on style embeddings Horvitz et al. (2024). The original system is trained on the Reddit Million User Dataset (MUD) (Khan et al., 2021) with a **four step** procedure:

1. A modified T5 model (Raffel et al., 2020) is trained to reconstruct texts from paraphrases and style embeddings (Wegmann).
2. This unsupervised model is then used to generate style transfer outputs between multiple authors from the original corpus.
3. The resulting synthetic data is then filtered using style embedding (Wegmann) distance and meaning preservation metrics.
4. Finally, a model is then fine-tuned on the resulting filtered dataset.

We reproduce the TinyStyler procedure with the exact dataset and hyperparameters in the original paper. We make only one modification: replacing Wegmann embeddings with STYLEDISTANCE in the generation and filtering steps. We include the formality transfer evaluation results in Table 12. In these automatic evaluations, the STYLEDISTANCE conditioned model performs comparably. We additionally include examples comparing model output in Table 13.

| Method                           | Acc ( $\rightarrow F, \rightarrow I$ ) | Sim ( $\rightarrow F, \rightarrow I$ ) | Fluency ( $\rightarrow F, \rightarrow I$ ) | Joint ( $\rightarrow F, \rightarrow I$ ) | GPT-2 |
|----------------------------------|----------------------------------------|----------------------------------------|--------------------------------------------|------------------------------------------|-------|
| TSTYLER <sub>WEGMANN</sub>       | 0.94 (0.90, 0.98)                      | 0.82 (0.81, 0.82)                      | 0.77 (0.83, 0.72)                          | 0.78 (0.77, 0.80)                        | 112.5 |
| TSTYLER <sub>STYLEDISTANCE</sub> | 0.96 (0.94, 0.98)                      | 0.80 (0.80, 0.81)                      | 0.77 (0.82, 0.73)                          | 0.80 (0.79, 0.80)                        | 112.4 |

Table 12: We reproduce the automatic formality transfer evaluation procedure from TinyStyler (Horvitz et al., 2024) on the GYAFD dataset (Rao and Tetreault, 2018).  $\rightarrow F$  corresponds to formal transfer, and  $\rightarrow I$  corresponds to informal transfer.

| Source Text (Informal)                                                                        | WEGMANN ( $\rightarrow$ Formal)                                 | STYLEDISTANCE ( $\rightarrow$ Formal)                           |
|-----------------------------------------------------------------------------------------------|-----------------------------------------------------------------|-----------------------------------------------------------------|
| <i>"its keeping me up at nite, i have to know what it is"</i>                                 | <i>"Have to know what this is, keeping me up at night."</i>     | <i>"What is it? It is keeping me up at night."</i>              |
| <i>"i like journey's open arms... but i like mariah carey's version of that song better."</i> | <i>"I prefer mariah carey's version of Journey, Open Arms."</i> | <i>"I think mariah carey's version of that song is better."</i> |
| <i>"And you can't rely on rumors."</i>                                                        | <i>"You can never trust rumors about it."</i>                   | <i>"I would say that you can't rely on rumors."</i>             |
| Source Text (Formal)                                                                          | WEGMANN ( $\rightarrow$ Informal)                               | STYLEDISTANCE ( $\rightarrow$ Informal)                         |
| <i>"I favor the man as he is humorous and grounded."</i>                                      | <i>"i love tahm as he is humorous and grounded..."</i>          | <i>"i'm gonna go with the man...hes humorous and grounded."</i> |
| <i>"I am sure you will both enjoy it."</i>                                                    | <i>"ok i am sure you both will enjoy it!"</i>                   | <i>"oh yea...you'll both enjoy it"</i>                          |
| <i>"I very much enjoy this song."</i>                                                         | <i>"i love this song so much, i'm a fan of it."</i>             | <i>"i like this song so much man..."</i>                        |

Table 13: Outputs from TinyStyler, conditioned on Wegmann embeddings and STYLEDISTANCE embeddings.

## J Resources

We provide links and citations to resources used in this paper which provide license information, documentation, and their intended use. Our usage follows the intended usage of all resources.

We utilize the following models:

- GPT-4 (full model, accessed June, 2024) (OpenAI et al., 2024)
- RoBERTa (roberta-base) (Liu et al., 2019; Devlin et al., 2019; Reimers and Gurevych, 2019)
- Learning Universal Authorship Representations (LUAR) Embedding model (Rivera-Soto et al., 2021)
- Style embedding model from Wegmann et al. (2022)
- LISA model (Patel et al., 2023)
- CoLA model (textattack/roberta-base-CoLA) (Warstadt et al., 2019)
- all-mpnet-base-v2 Sentence Similarity model (Reimers and Gurevych, 2019) (<https://huggingface.co/sentence-transformers/all-mpnet-base-v2>)

We utilize the following datasets and resources:

- STEL dataset (Wegmann and Nguyen, 2021)
- Contrastive Authorship Verification dataset (Wegmann et al., 2022)
- C4 (Raffel et al., 2020)
- LIWC (Tausczik and Pennebaker, 2010)
- PAN 2011 - 2015 Authorship Verification Datasets (Argamon and Juola, 2011; Juola and Stamatatos, 2013; Stamatatos et al., 2014, 2015) - Curated at: <https://huggingface.co/datasets/swan07/authorship-verification>
- GYAFC (Rao and Tetreault, 2018)

We utilize the following software:

- DataDreamer (Patel et al., 2024)
- Transformers (Wolf et al., 2019)
- Datasets (Lhoest et al., 2021)
- PEFT (Mangrulkar et al., 2022)
- Sentence-Transformers (Reimers and Gurevych, 2019)
- NLTK (Bird and Loper, 2004)

We estimate the total compute budget and detail computing infrastructure used to run the computational experiments found in this paper below:

- 8x NVIDIA RTX A6000 / 100GB RAM / 16x CPU – 80 hours