Style transfer with large language models

Guillem Chillon[1], Jordan Matelsky[2], and Zixuan Bian[3]

[1]guillemc@seas.upenn.edu
[2]matelsky@upenn.edu
[3]bianzx@seas.upenn.edu

May 2025

Abstract

Style transfer in text generation is a challenging problem due to the difficulty of altering style independently of content. Achieving this requires manipulating text in a way that changes its style while preserving its original meaning. We propose fine-tuning a compact language model using disentanglement techniques, specifically employing learning objectives that incorporate stylistic awareness. Our approach aims to improve neural network's ability to perform style transfer. The expected outcome is enhanced style transfer performance with a focus on achieving high style accuracy and maintaining content integrity. Successfully implementing this method would significantly enhance controllable text generation. This advancement has potential applications in both industrial settings, like personalized marketing; human communication, enabling clarity and disambiguation; and academic research, potentially expanding the capabilities of natural language processing in generating text that meets specific stylistic criteria.

1 Motivation

The rapid advancement of large language models (LLMs) has revolutionized natural language processing (NLP). Although large language models (LLMs) have been used anecdotally for style transfer, they typically do not explicitly disentangle style from content, leading to possible unintended alterations in meaning. Our project aims to bridge this gap by investigating the layers in which style-specific attributes are most effectively represented, allowing independent manipulation of style and content.

2 Related Work

Early approaches to text style transfer aimed to disentangle style, drawing inspiration from analogous techniques in image style transfer [Gatys et al., 2016, Zhu et al., 2017] and subsequently adapted for NLP applications [John et al., 2019, Patel et al., 2022]. These methods demonstrated that fine-tuned pretrained LMs can effectively induce stylistic transformations even with non-parallel data. However, these approaches often treat style and content as inherently intertwined, making it challenging to isolate and control individual attributes.

Recent research on transformer models has advanced our understanding of how semantic information is organized across layers. In particular, work on scaling monosemanticity [Templeton et al., 2024] reveals that individual neurons or directions in transformers often capture single, well-defined semantic features. This finding spurs further investigations into how specific linguistic attributes, including style-related properties, are encoded in deep networks.

Building on it, our work adopts a more targeted strategy. By applying linear probes, the present study directly measures how formality and domain-specific features are localized across the layers of an LLM. Such an approach offers the potential to design layer-specific interventions that enhance style transfer performance while preserving content integrity.

3 Datasets

We propose using style- and domain-aware datasets, such as:

- *Pavlick Formality Dataset*: Contains formality scores for a total of 11,274 text samples aggregated from different domains: Yahoo! Answers (4,977), blogs (1,821), emails (1,701), and news (2,775). Each record includes an `avg_score` (ranging from -3 to 3, where lower scores indicate less formal sentences) and the `sentence` itself.

- *Generated Formality Dataset*: Produced using GPT-3.5-turbo, this dataset comprises 500 examples in total, including 250 paired examples where each pair consists of an informal sentence and its corresponding formal version.

These datasets are suitable for our project due to their size and the nature of the style transfer tasks. Additionally, we will also evaluate the ability to maintain content for downstream tasks, such as:

- *Natural Questions*: Question answering (QA) dataset of +300K QA pairs. Each example has query text, and long- and short-forms answers.

- *CNN/Daily Mail*: Summarization dataset of ≈300K articles paired with human-written summaries.

This evaluation will ensure that while we modify style, the underlying semantic information remains intact.

4 Problem Formulation

The problem is formulated as a classification task where the goal is to predict the formality or domain label of a text sample based on its representation in an LLM. Linear probes are used to identify the layers where these attributes are most effectively encoded.

Let X represent the input text samples and Y the corresponding labels (formality or domain). The task is to learn a mapping $f : X \to Y$ using representations H_l extracted from layer l of the LLM. The classifier g is trained to minimize the loss function:

$$\mathcal{L}(g) = \frac{1}{N} \sum_{i=1}^{N} \ell(g(H_l(x_i))y_i) \tag{1}$$

where ℓ is the cross-entropy loss, N is the number of samples, and x_i, y_i are the input-label pairs.

5 Methods

Our methodology involves the following steps:

1. **Dataset Preparation:** Text samples are collected from Pavlick Formality Scores dataset. Text samples are preprocessed and split into training and testing sets. In addition, data augmentation strategies are employed to create paired formal-informal examples and avoid content discrepancies.

2. **Feature Extraction:** The Phi-4 language model is used to extract representations from different layers by processing text samples. state outputs from each layer are recorded to capture a hierarchy of linguistic features.

3. **Linear Probing:** Logistic regression classifiers are trained on the extracted features to predict formality and domain labels.

4. **Evaluation:** Evaluation is conducted on two fronts:
   Style Classification: The accuracy of the classifiers is measured on both the training and testing sets.

Content Preservation: Two downstream tasks assess how well the underlying semantic content is maintained after style modifications:

- For the question answering task using the Natural Questions dataset, content preservation is evaluated by the **F1 score**. In this context, Precision is defined as the ratio of correctly predicted answer tokens to the total predicted tokens, while Recall is the ratio of correctly predicted tokens to the total tokens in the reference answer. The F1 score, as the harmonic mean of precision and recall, captures the balance between these measures.
- For the summarization task using the CNN/Daily Mail dataset, content preservation is measured using the **ROUGE-1 metric**. ROUGE-1 computes the unigram overlap between the generated summary and the reference summary, indicating the degree of content similarity.

The evaluation metrics are formally defined as follows:

- **F1 Score (Natural Questions):**

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

- **ROUGE-1 (CNN/Daily Mail):**

$$\text{ROUGE-1} = \frac{\sum_{w \in \text{Ref}} \min\big(\text{Count}_{\text{match}}(w), \text{Count}_{\text{cand}}(w)\big)}{\sum_{w \in \text{Ref}} \text{Count}(w)} \tag{5}$$

These metrics ensure that, while stylistic attributes are accurately classified, the semantic content remains preserved following style modifications.

6 Experiments and Results

We conducted several experiments to probe the Phi-4 language model's capacity to encode stylistic attributes. Linear probes are employed to assess how well different layers represent formality and domain-specific features.

6.1 Formality Probing

Figure 1 displays the accuracy plot for formality probing on the Pavlick Formality Dataset. The results indicate that the intermediate layers (layer 13) of the model achieve the highest accuracy, suggesting that these layers are most effective at capturing formality-related features.

Figure 2 presents the accuracy plot for paired formality probing using the Generated Formality Dataset. In this experiment, paired formal-informal examples are used to assess the model's capability to distinguish between the two styles. Consistent with the single-sample analysis, the intermediate layers (layer 17) again demonstrate superior performance.

6.2 Domain Probing

Figure 3 shows the accuracy plot for domain probing on the Pavlick Formality Dataset. The findings reveal that earlier layers (layer 8) exhibit higher predictive accuracy for domain-specific attributes, further confirming that stylistic and domain information are localized in particular regions of the model.

6.3 Custom Model Training with Style Latent

Building upon our findings from the linear probing experiments, we implemented a custom model that incorporates a style latent representation. We modified the architecture of the Phi-4 model by introducing a style embedding layer at layer 13 for formality control and layer 8 for domain-specific features, as these were identified as reasonable insertion layers in our probing experiments (i.e., well past the "burn-

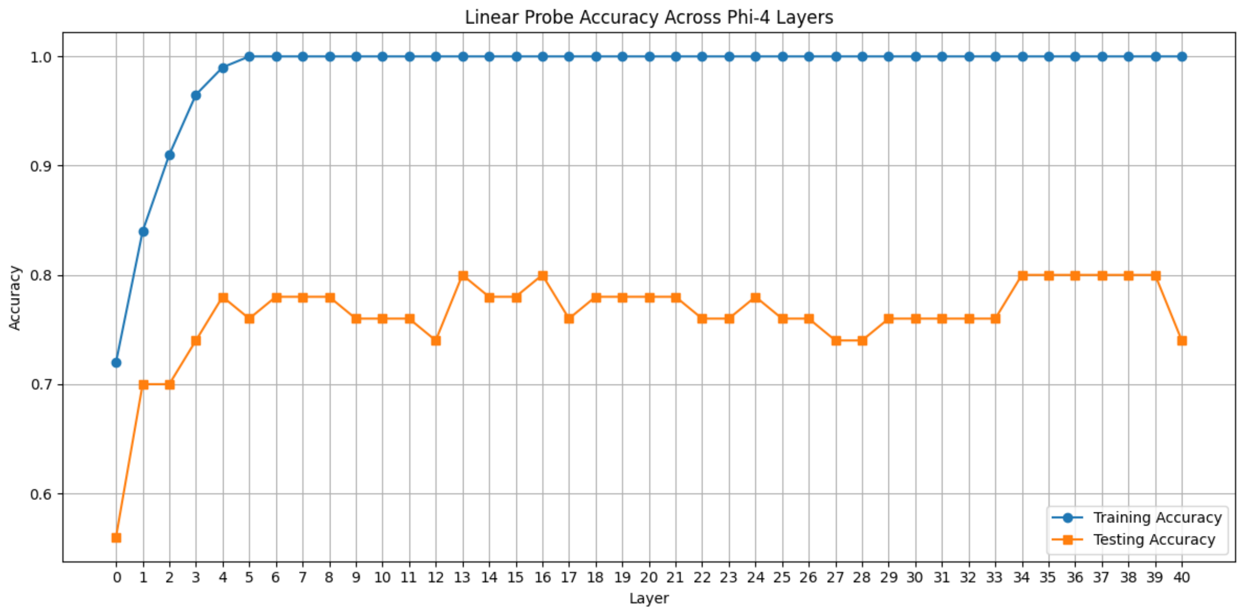

Figure 1: Formality Probing Accuracy Across Phi-4 Layers

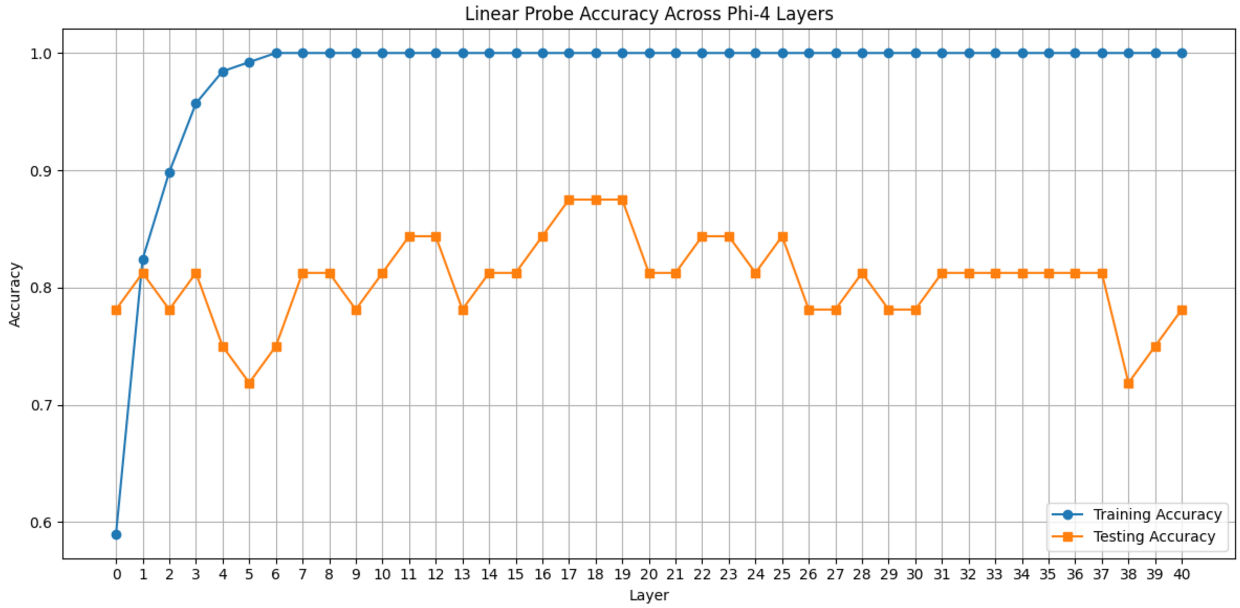

Figure 2: Paired Formality Probing Accuracy Across Phi-4 Layers

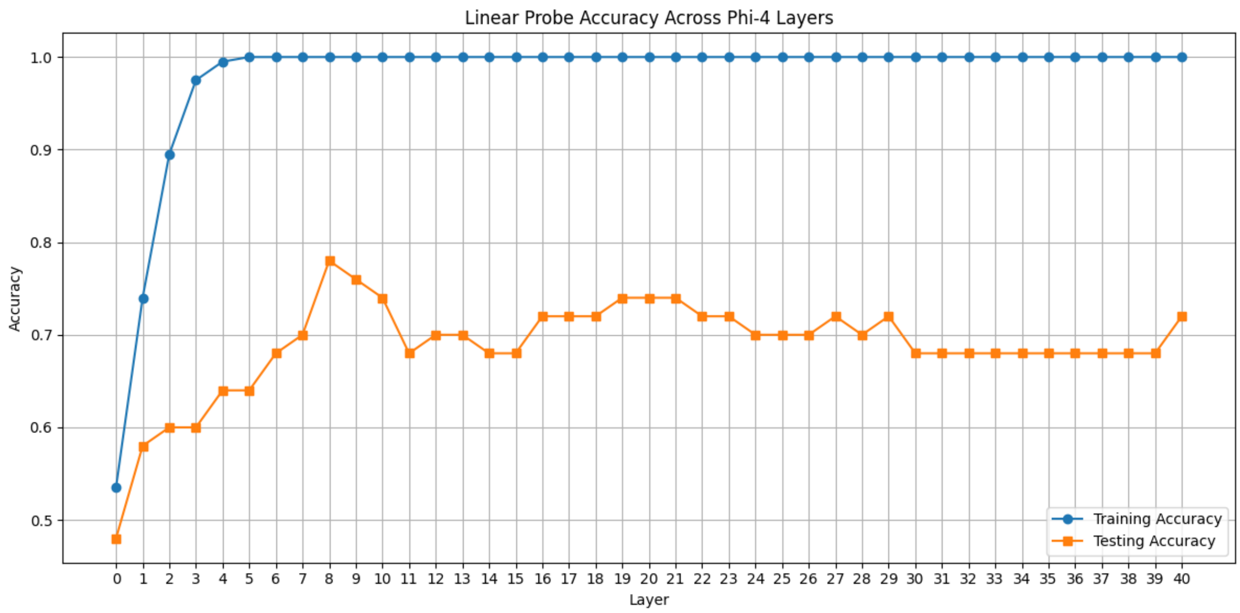

Figure 3: Domain Probing Accuracy Across Phi-4 Layers

in" phase of the early model, relatively close to the center of the stack of attention modules).

6.3.1 Style Embedding Architecture

The style embedding layer consists of a dimension reduction component that maps the original hidden dimension (2560) to a smaller latent space (n=8), followed by a projection back to the original dimension. This structure allows the model to learn a compressed representation that effectively captures style information while preserving content.

$$h_{style} = W_{proj} \cdot \text{ReLU}(W_{reduce} \cdot h_{original} + b_{reduce}) + b_{proj} \quad (6)$$

During training, we froze the parameters of the base model and only updated the style embedding layer parameters, ensuring that content representations remained stable while style-specific parameters were optimized. This meant that the layer served effectively as an "adapter" layer, serving to change the dimensionality of the hidden state without — we hoped — altering the underlying content representation.

6.3.2 Training Process

The model was trained on the Pavlick Formality Dataset for 20 epochs (Fig. 5) using the Adam optimizer with a learning rate of 1e-4. We employed a weighted cross-entropy loss to account for class imbalance:

$$\mathcal{L}_{style} = -\sum_{i=1}^{N} w_i \cdot y_i \log(\hat{y}_i) \quad (7)$$

where w_i represents the class weight for sample i, y_i is the true label, and $\hat{y}_i$ is the predicted probability.

6.3.3 Domain Classification Performance

Figure 4 presents the confusion matrix for domain classification using our style latent approach. The model achieves high accuracy in distinguishing between different domains, with particularly strong

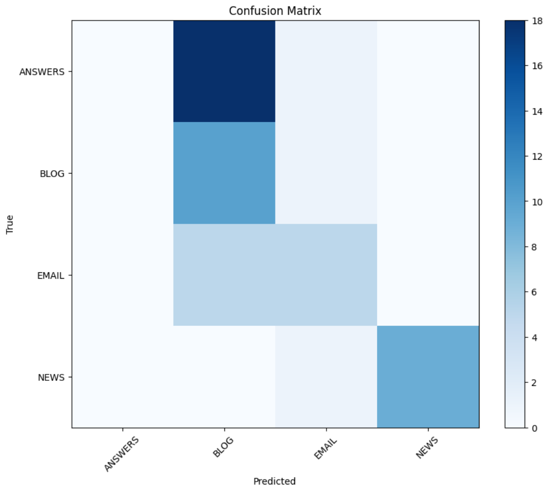

Figure 4: Confusion Matrix for Domain Classification using the Style Latent Model

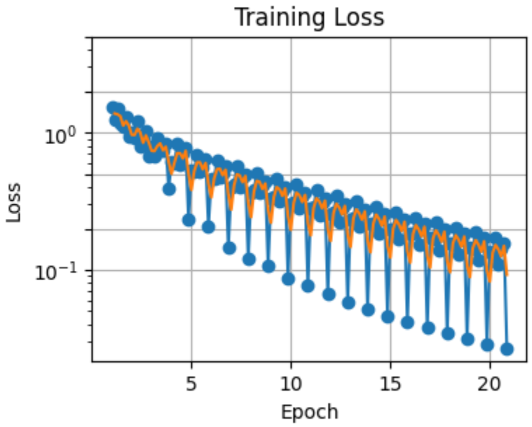

Figure 5: Training Loss Curve for Style Latent Model.

performance in identifying content from the EMAILS and NEWS categories.

The confusion matrix reveals that the model occasionally confuses EMAIL with BLOG content, which is reasonable given their potential stylistic similarities. However, it maintains clear separation between formal domains (NEWS) and informal domains (ANSWERS), demonstrating the model's ability to capture meaningful stylistic distinctions.

6.3.4 Content Preservation

Content preservation metrics were integrated by evaluating the Natural Questions dataset using the F1 score for question answering, yielding an F1 score of 0.72. Additionally, we evaluated both the baseline and style-modified models on the CNN/Daily Mail dataset using the ROUGE metric for summarization, as shown in Table 1.

These results confirm no significant compromise in semantic integrity despite style modifications, with only a small reduction (approximately 3%) in ROUGE scores compared to the baseline model. This minimal degradation in content preservation metrics demonstrates that our style embedding approach effectively separates style representation from content, allowing style manipulation without substantial impact on the underlying meaning.

7 Conclusion and Future Work

Our research demonstrates that style attributes can be effectively isolated in specific layers of transformer-based language models. Using linear probing techniques, we identified that formality features are predominantly encoded in intermediate layers, while domain-specific features are captured in earlier layers. This localization enables targeted interventions that can manipulate style while preserving content.

The style latent approach introduced in this paper shows promise for controllable text generation, allowing for explicit style transfer with minimal impact on semantic content. Future work could explore more granular style attributes beyond formality and domain, as well as investigate methods to further improve the disentanglement of style and content representations.

References

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks.

Table 1: ROUGE Score Comparison Between Baseline and Style-Modified Phi-4 Models

Model	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**ROUGE-Lsum**
Baseline Phi-4	0.1518	0.0776	0.1071	0.1253
Modified Phi-4	0.1470	0.0753	0.1049	0.1224
Difference	-0.031	-0.029	-0.020	-0.023

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. Disentangled representation learning for non-parallel text style transfer. pages 424–434, July 2019. doi: 10.18653/v1/P19-1041. URL `https://aclanthology.org/P19-1041/`.

Ajay Patel, Nicholas Andrews, and Chris Callison-Burch. Low-resource authorship style transfer: Can non-famous authors be imitated? *arXiv preprint arXiv:2212.08986*, 2022.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, 2024. URL `https://transformer-circuits.pub/2024/scaling-monosemanticity/`. Accessed: 2024-04-07.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017.