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 finetuning 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. Furthermore, we aim to modify these layers to determine if tweaking them can produce results that align more closely with the desired style.

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 LLMs 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 this, our work adopts a more targeted strategy. By applying linear probes, we directly measure how formality and domain-specific features are localized across the layers of an LLM. Based on these findings, we design layer-specific interventions targeting the most style-aware sections of the model to 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 with:

- *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 to 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: To assess how well the underlying semantic content is maintained after style modifications, content preservation is measured using:

- **BLEU**: BLEU-1, BLEU-2, BLEU-3, and BLEU-4 are used to evaluate the n-gram overlap between the generated summary and the reference summary. These metrics provide a quantitative measure of how well the generated text aligns with the original content.
- **ROUGE**: ROUGE-N computes the overlap between the generated summary and the reference summary, indicating the degree of content similarity. ROUGE-L evaluates the length of the longest common subsequence (LCS) between the candidate and reference, emphasizing sentence-level structure and fluency. ROUGE-Lsum extends ROUGE-L by computing LCS-based similarity across sentences and aggregating the result, offering a more robust evaluation for multi-sentence or document-level summaries.

The evaluation metrics are formally defined as follows:

- **BLEU-N:**

$$\text{BLEU-N} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \log p_n\right) \tag{2}$$

Where BP is the brevity penalty, p_n is the precision of n-grams, and w_n is the weight for each n-gram.

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases} \tag{3}$$

Here, c is the length of the candidate summary, and r is the reference length.

- **ROUGE-N:**

$$\frac{\sum_{g \in \text{Ref}_N} \min\left(\text{Count}_{\text{match}}(g), \text{Count}_{\text{cand}}(g)\right)}{\sum_{g \in \text{Ref}_N} \text{Count}(g)} \tag{4}$$

Where g is an n-gram and Ref_N denotes all reference n-grams of order N.

- **ROUGE-L:**

$$R_{LCS} = \frac{\text{LCS}(X,Y)}{\text{length}(Y)}, \quad P_{LCS} = \frac{\text{LCS}(X,Y)}{\text{length}(X)} \tag{5}$$

$$\text{ROUGE-L} = \frac{(1+\beta^2) \cdot R_{LCS} \cdot P_{LCS}}{R_{LCS} + \beta^2 \cdot P_{LCS}} \tag{6}$$

Where $\text{LCS}(X,Y)$ is the length of the longest common subsequence between candidate X and reference Y. $\beta = 1$ to equally weight precision and recall.

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 domain-specific attributes are represented across all layers of the model, with no single layer being exclusively responsible for selectively encoding domain information. It further confirms that stylistic and domain information are not localized in particular layers but are distributed across the model.

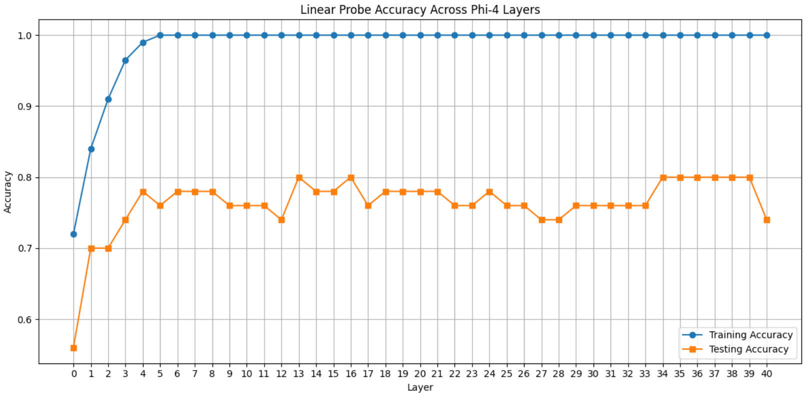

Figure 1: Formality Probing Accuracy Across Phi-4 Layers

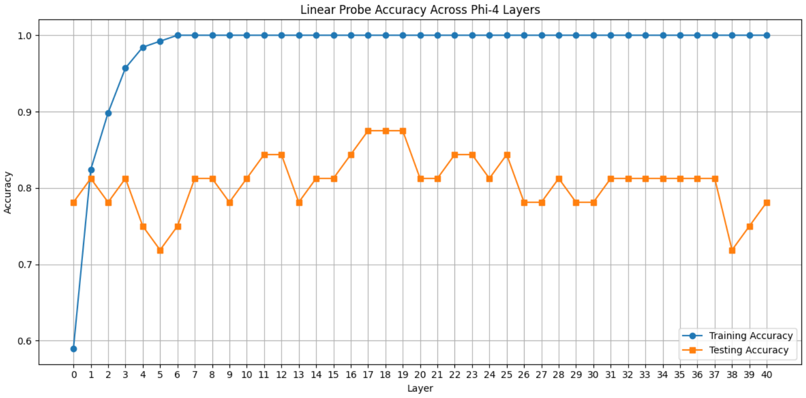

Figure 2: Paired Formality Probing Accuracy Across Phi-4 Layers

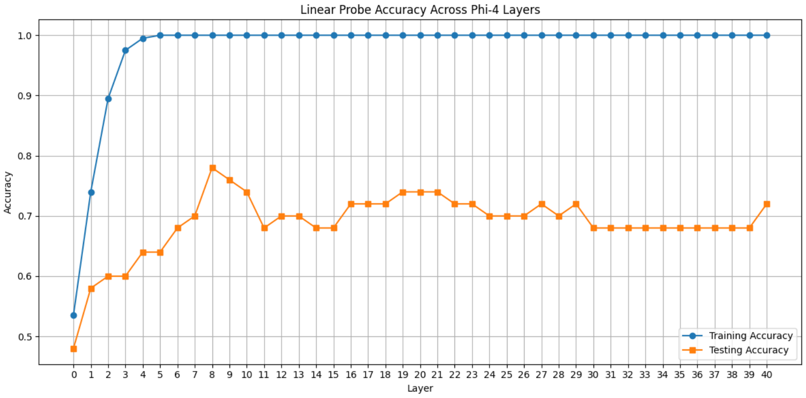

Figure 3: Domain Probing Accuracy Across Phi-4 Layers

Table 1: Content Preservation Metrics of Baseline and Style-Modified Model

Metric	Baseline	Modified	Diff.
BLEU-1	0.0767	0.0761	-0.0006
BLEU-2	0.0503	0.0501	-0.0002
BLEU-3	0.0353	0.0352	-0.0001
BLEU-4	0.0249	0.0248	-0.0001
ROUGE-1 F1	0.1584	0.1582	-0.0002
ROUGE-2 F1	0.0796	0.0795	-0.0001
ROUGE-L F1	0.1113	0.1111	-0.0001
ROUGE-1 Precision	0.0888	0.0886	-0.0002
ROUGE-1 Recall	0.8897	0.8897	0.0000

Table 2: Examples of Style Transfer with Different Domain Latents

Style	Generated Text
Original	The MDC has a long list of complaints it alleges are non-fulfillment of the Global Political Agreement and wants resolved before it can participate fully in government.
ANSWERS	MDC also accused the ZANU PF government of being a one party state and said the MDC is not treated as an equal partner...
BLOG	So, MDC accused the government of failing to address key issues.
EMAIL	the MDC-T says it wants the reserve bank gov to be replaced and that it wants to know what is in the arms deal signed by the government?
Original	How does it affect dogs?
ANSWERS	What are the symptoms? What should I do if my dog has it? And, can my dog give it to me?
BLOG	Dogs can suffer from various neurological disorders, including epilepsy, which can cause seizures and other symptoms affecting their nervous system.
NEWS	The disease can cause severe symptoms in dogs, including diarrhea, vomiting, and lethargy. It can be particularly dangerous for puppies, elderly dogs, or those with compromised immune systems.
EMAIL	**What are common signs of poisoning in dogs?** - **Answer:** Common signs include vomiting, diarrhea, excessive drooling, lethargy, difficulty breathing, seizures, and loss of coordination.

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-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.

$$h_{style} = W_{proj} \cdot \text{ReLU}(W_{reduce} \cdot h_{original} + b_{reduce}) + b_{proj} \tag{7}$$

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. 4) using the Adam optimizer with a learning rate of 1e-4.

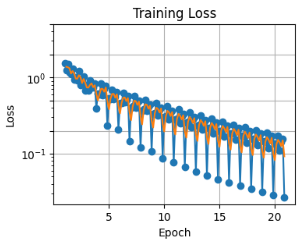

Figure 4: Training Loss Curve for Style Latent Model.

We employed a weighted cross-entropy loss to account for class imbalance:

$$\mathcal{L}_{style} = -\sum_{i=1}^{N} w_i \cdot y_i \log(\hat{y}_i) \tag{8}$$

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 5 presents the confusion matrix for domain classification using our style latent approach. The model achieves *accuracy* = 0.54, being able to distinguish between different domains, with particularly strong performance in identifying content from the EMAILS and NEWS categories.

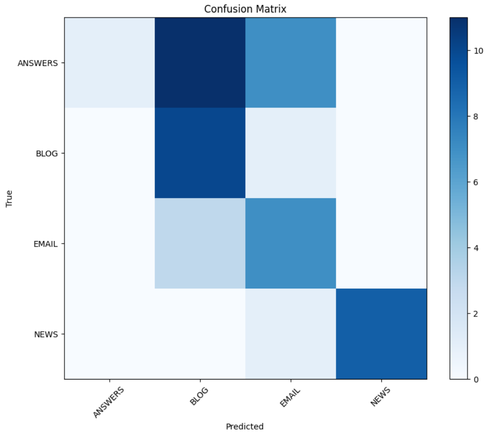

Figure 5: Confusion Matrix for Domain Classification using the Style Latent Model

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 both the baseline and style-modified models on the CNN/Daily Mail dataset, as shown in Table 1.

A negligible reduction ($< 0.01\%$) in BLUE and ROUGE scores was observed for the style-modified model compared to the baseline, confirming no compromise in semantic integrity despite style modifications. 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.

6.3.5 Style Transfer Examples

Table 2 illustrates examples of style transfer using our approach, where we applied different domain-specific latent vectors to the same content. The table showcases how the original text is rendered differently when encoded with ANSWERS, BLOG, and EMAIL style latents. Note how each variant maintains similar (but clearly imperfect) core information while adopting characteristic stylistic elements of its domain: the ANSWERS style uses more accusatory language, the BLOG style incorporates temporal markers and detailed reporting, and the EMAIL style employs a more direct, abbreviated approach with question formatting. These examples demonstrate the model's ability to transform text style while preserving the underlying semantic content. In the second example, notice the confusion where the EMAIL generation (which begins with *What are common signs of poisoning in dogs?*) begins with a question and forms what appears to be an ANSWERS style response, reflecting the findings from the domain probing experiments that EMAIL and ANSWERS are stylistically similar and often confused (Fig. 5).

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. *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.