

Dynamic Multi-Reward Weighting for Multi-Style Controllable Generation

Karin de Langis and Ryan Koo and Dongyeop Kang
University of Minnesota
{dento019, koo00017, dongyeop}@umn.edu

Abstract

Textual style expresses a diverse set of information, including interpersonal dynamics (e.g., formality) and the author’s emotions or attitudes (e.g., disgust). An open question is how language models can be explicitly controlled so that they weave together target styles when generating text: for example, to produce text that is both negative and non-toxic. One approach to such controlled generation is multi-objective reinforcement learning (RL), but how to best combine multiple objectives in a reward function is an open question. In this paper, we investigate various formulations of multi-style rewards, including calibrated outputs from discriminators and dynamic weighting by discriminator gradient magnitudes. We find that our proposed dynamic weighting outperforms static weighting approaches with respect style control while maintaining linguistic quality, and we explore its effectiveness in 2- and 3-style control. All code and data for the RL pipelines will be publicly available.¹

1 Introduction

Textual style is an important component of communication that conveys information not included in the text’s raw semantic content (Hovy, 1995). Consequently, it is vital that language models can understand and apply styles themselves. Prior work has explored the domain of controlled style generation, a task in which a generative language model aims to generate text with a specified style². However in practice, text frequently contains not only a single style, but a combination of styles (Kang and Hovy, 2021). For example, consider being asked to give feedback to a colleague at work: both formal and positive styles would be appropriate. On the other hand, if speaking with a friend about a

¹<https://github.com/minnesotanlp/dynamic-multi-reward-weighting>

²Style transfer, on the other hand, involves paraphrasing given text to a target style without altering its original meaning

Prompt: The satire is just...

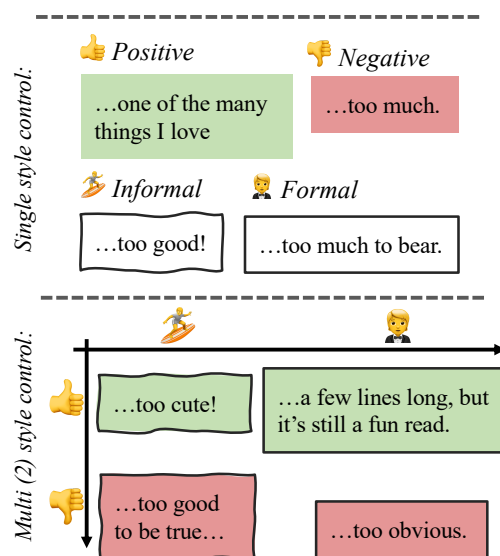


Figure 1: An example of our 1- and 2-style models generating completions to a given prompt. Models are trained with reinforcement learning where the reward is derived from the target styles’ discriminators.

movie, both informal and positive styles are likely to be useful. Examples of generations following multiple styles models can be found in Figure 1.

Especially as large language models (LLMs) grow in capability and popularity, it is desirable to include fine-grained control of the styles in LLM outputs. For instance, in almost all cases, toxicity and hate speech must be tightly controlled such that the model does not produce harmful output. At the same time, in response to the user’s preferences or the application, it can be beneficial for the LLM to simultaneously control additional attributes such as humor, formality, or the use of figurative language. In order to achieve these goals reliably, techniques for robust multi-style control are needed.³

³While some sociolinguistics theories distinguish between textual style and textual attributes, in this work, we follow the common convention in recent NLP papers of broadly using ‘style’ to encompass both of these ideas (Jin et al., 2022).

Controlling for more than one style during generation is an under-investigated area, with prior work focusing on controlling for a single style, or a style and a target topic(s) (Keskar et al., 2019; Liu et al., 2022). In this work, we investigate the use of Reinforcement learning (RL) for controlling multiple styles. RL approaches satisfy multiple desiderata for generations by employing a reward function in which each individual desideratum contributes to the reward; this approach is recently gaining more interest in the alignment literature (see e.g., fine-grained reinforcement learning from human feedback (RLHF) from Wu et al. (2023)). In this work, we apply a similar approach for multi-style controlled generation, in which “style scores” from individual style discriminators are combined into a single reward function during RL fine-tuning.

The optimal approach to combining multiple reward signals into a reward function is an open question. To further explore this question, we implement several strategies for formulating the multi-style reward, including a novel dynamic-weighting approach. Interestingly, our evaluations indicate that dynamically weighting each component outperforms static weighting, and we also find that simple steps such as confidence calibration and binarization of style discriminator output can improve model performance. We also implement a custom plug-and play pipeline (Dathathri et al., 2020) for comparison.

This is to our knowledge a first-of-its-kind work investigating multi-style controllable generation through an RL lens with a new reward shaping approach via dynamic gradient-based weighting. Work on multi-reward formulation for RL is especially relevant in the current landscape, given the modern alignment techniques that incorporate multiple axes of human feedback, as recent work investigates composing multiple types of human feedback – e.g., helpfulness, correctness – into the reward function (Rame et al., 2023).

2 Related Work

Controlled Text Generation Methods for controlled generation can generally be grouped into three main categories: fine-tuning, retraining, and post-processing (Zhang et al., 2023). Post-processing approaches are the most lightweight and involve applying transformations during decoding, rather than making any adjustments to model weights themselves. Examples of such methods

include plug and play, or PPLM (Dathathri et al., 2020)), which uses gradients from an attribute classifier to guide the language model’s hidden state; generative discriminators (GeDI) which compute control codes and anti-control codes for all possible next tokens (Krause et al., 2020); and Attribute Alignment, which learns an alignment function (Yu et al., 2021) infuse attribute representations into a pre-trained language model to guide generation. Prefix-tuning (Li and Liang, 2021; Qian et al., 2022) can also guide generation by prepending task- or style-specific “prefix” vectors. Retraining (or refactoring) methods involve retraining language models from the ground up on the control task; for example, CTRL (Keskar et al., 2019) retrains a class-conditional language model conditioned on many control codes to guide generations. Another retraining approach is Cev-LM (Moorjani et al., 2024), a prototype-then-edit semi-autoregressive language model that applies edit vectors in the latent space.

Our work falls under the fine-tuning category. Fine-tuning methods adjust parameters of a pre-trained LLM toward fulfilling the desired controls. Reinforcement learning (RL) is a common fine-tuning approach for controlled text generation (Zhang et al., 2023), e.g., Gong et al. (2019) use a style classifier model to provide a target style reward, and Upadhyay et al. (2022) use token-level dense rewards and taking the weighted sum of these rewards that were heuristically determined to update the policy. Other works align models towards specific attributes by modeling the reward function as a preference model (Rafailov et al., 2023) to bypass the need for explicitly calculating a reward function, reducing the task to a maximum likelihood objective. More recently, various approaches to fine-tuning language models via human preferences (Ziegler et al., 2020; Ouyang et al., 2022) have seen success in guiding text generations to be more aligned with desired attributes.

Finally, we point out that textual style transfer is related to controlled generation, but it is a distinct task that involves transforming an input text’s style while preserving the semantics. Recent work in style transfer include Steering Vectors (Subramani et al., 2022), which inject steering vectors into the model during decoding. Variations such as ActAdd (Turner et al., 2023) and activation engineering (Konen et al., 2024) have also been proposed.




Approach	Calculation	Brief rationale
 Confidence Calibration	$\frac{d_i(x)}{\text{temperature}}$	Temperature scaling for rewards that better reflect model confidence
 Binarized	$\begin{cases} -1, & d_i(x) < 0.5 \\ 1, & d_i(x) \geq 0.5 \end{cases}$	Simplifies reward process: model confidence plays no role
 Dynamic Weighting	$w(1 - d_i(x))$	Pick which attribute should contribute most to reward based on performance

Figure 2: Three different approaches to combine multiple rewards effectively for reinforcement learning. $d_i(x)$ refers to the attribute from discriminator i on input text x . We find dynamic weighting (our proposal) is particularly effective for multi-style control.

Multi-Objective Rewards Recent work explores how to incorporate reward signals from multiple sources into reward functions, particularly in alignment literature involving RLHF. Notably, recent work investigates multi-objective RLHF by training separate reward models from human preference data and linearly combining those rewards (Wu et al., 2023; Rame et al., 2024; Ramé et al., 2024). In this work, we apply a similar approach for controlled style generation, and we also experiment with additional approaches to reward combination.

Inverse RL Our work in dynamically shaping the weights of the reward function is also adjacent to Inverse Reinforcement Learning (IRL), which learns both a policy and a reward function. IRL alternates policy update steps with reward model update steps, requiring expert demonstrations for the reward update step. For example, Ghosh et al. (2021) use IRL for Table-to-Text generation, and their reward function takes several table descriptors as inputs; and Fu et al. (2023) use IRL for the text summarization task, with a reward function that combines various sub-rewards like saliency and coverage. However, we clarify a distinction between our work and IRL: whereas IRL explicitly learns a reward function, our dynamic weighting method is best characterized as reward shaping for online policy update. Specifically, we do not use expert demonstrations to optimize the weights of the reward function, but instead dynamically "shape" the coefficients of the fine-grained reward function components during the policy update step by inspecting the gradients for each component.

3 Proposed Method: Multi-Reward Control in RL

In this section we provide a preliminary formulation of using reinforcement learning (RL) for a

single style control and then formulate the specific multi-reward control functions we propose.

3.1 Preliminary Formulation: Style Control using Reinforcement Learning

Reinforcement learning for language models frames the generative language model as a *policy* network, π_θ . The policy is a probability distribution over all possible actions (i.e., vocabulary tokens) that determines a_t , i.e., the action to be taken at timestep t given the state x_{t-1} . We can then generate tokens based on $\pi_\theta(a_t|x_{t-1})$. Reinforcement learning also introduces a reward function R that takes in a state x_t and outputs a scalar valuation of that state. (Reinforcement learning for language models generally uses sparse rewards, i.e., a reward is only computed once the full sequence of tokens have been generated.) The objective of reinforcement learning, then, is to learn a new policy $\pi_{\theta'}$ such that the policy maximizes the expected value of R . Controlling a single style using RL fine-tuning typically formulates the reward using a discriminator d for the target style, as in $R(x) = d(x)$.

3.2 Multi-Reward Control Formulations

When constructing a reward formulation that combines outputs from multiple style discriminators, we consider two important aspects. The first is the discriminator output itself: we want to find a way to express the output, e.g., via some transformation, that provides a strong and consistent reward signal.

The second aspect we consider is the how to weigh the combined reward. This is important because training can become intractable if the signal from one discriminator dominates. For instance, consider a 2-style softmax reward formulation, and imagine our starting policy is such that there is a near-zero likelihood of producing a generation that results in high softmax scores for both d_1 and

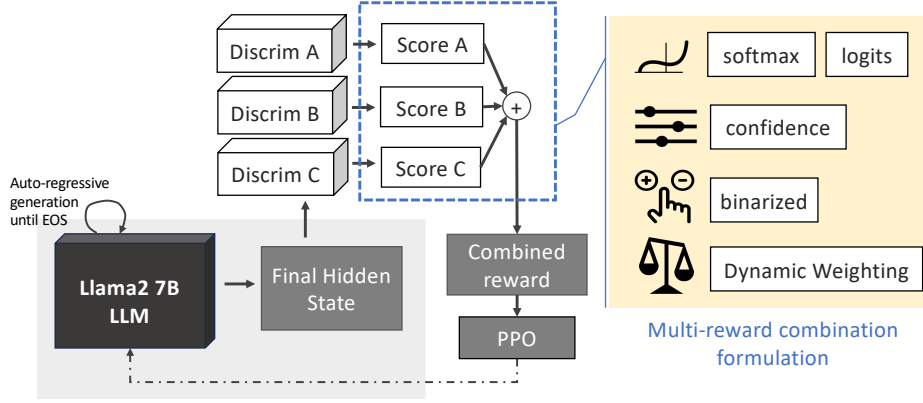


Figure 3: Pipeline for three-style control using fine-tuning Llama 7B model with Proximal Policy Optimization (PPO). We investigate several techniques for integrating feedback from style discriminators into the reward.

d_2 , but there is a high likelihood of producing generations that results in a high score for d_1 . The policy will quickly discover how to move in a direction that maximizes scores from d_1 and move away from exploring the states that result in more balanced rewards. Instead, we want to combine the discriminators in a way that encourages balancing all outputs.

Motivated by these considerations, we explore multiple approaches to calculate a reward R for a generation x by combining output from the attribute discriminators d_1, d_2, \dots, d_n with corresponding target styles k_1, k_2, \dots, k_n , and we write the logit value for the target class as $d_i(x)_{k_i}$. We consider the following multi-reward formulations:

Logits We take the logits of the target class from the discriminator output, $R = \sum_{i=1}^n d_i(x)_{k_i}$.

Softmax We take the softmax σ values of the target class from the discriminator output, $R = \sum_{i=1}^n \sigma(d_i(x))_{k_i}$.

Binarized When discriminators are unsure of their prediction, the reward signal becomes noisier, potentially hampering the policy learning process. Thus, we consider binarized rewards to make the signals more discrete and emphasized.

$$R = \sum_{i=1}^n w_i, \quad w_i = \begin{cases} 1 & \sigma(d_i(x))_{k_i} \geq 0.5 \\ -1 & \text{otherwise} \end{cases} \quad (1)$$

Calibrated Logit scores for are adjusted following Guo et al. (2017) in order to calibrate model confidence. Similar to the binarized reward, the calibrated reward helps better calibrate the low confidence in the discriminators' predictions to find a

more accurate policy.

$$R = \sum_{i=1}^n \frac{d_i(x)_{k_i}}{\text{temperature}_i} \quad (2)$$

Dynamic Weighting We calculate a dynamic weight w_i . The motivation for the dynamic weighting approach is that when combining multiple style discriminators, it is not always clear whether a signal from one discriminator should be prioritized over another. Our approach weighs the result from each discriminator by considering the magnitude of the gradient of the cross entropy loss for $d_i(x)$ with respect to the desired style. Then, our reward function is

$$R = \sum_{i=1}^n w(1 - \sigma(d_i(x))_{k_i}) \quad (3)$$

$$w = \begin{cases} \text{grad_norm}_i & \text{if } \sigma(d_i(x))_{k_i} > 0.5 \\ -\text{grad_norm}_i & \text{otherwise} \end{cases} \quad (4)$$

where grad_norm_i is the normalized magnitude of $d_i(x)$'s gradient:

$$\text{grad_norm}_i = \frac{\|\nabla_{d_i(x)} \mathcal{L}_{CE}\|}{\sum_{j=0}^n \|\nabla_{d_j(x)} \mathcal{L}_{CE}\|} \quad (5)$$

We compose all reward shaping methods by taking the convex combination of each discriminator output, with only the gradient scaling method differing. For example, the binarized, confidence, and calibration methods are all follow the formulation with given d_1 and d_2 (where each d_n refers to some style classifier)

$$R_{\text{final}} = \alpha \cdot d_1(x) + (1 - \alpha) \cdot d_2(x) \quad (6)$$

where each d_1 and d_2 rewards are shaped via each respective method. Here we take the α as $1/n$

	Sentiment	Formality	Irony	Emotion (Eckman Seven)	Toxicity
Labels	positive, negative	formal, informal	ironic, not ironic	Fear, anger, joy, sadness, dis- gust, surprise, neutral	toxic, non-toxic
Dataset	SST2	GYAFC	SemEval	go-emotions	Jigsaw
F1 (Macro)	0.90	0.90	0.78	0.53	0.80

Table 1: We train discriminators on each of the above styles (columns).

where n is the number of discriminators and can be extended to any number n such that $\sum_{i=1}^n \alpha_i = 1$. For the gradient-weight scaling, we replace α with grad_norm_i (extendable to any n number of discriminators)

$$R_{\text{final}} = \text{grad_norm}_1 \cdot d_1(x) + \text{grad_norm}_2 \cdot d_2(x) \quad (7)$$

4 Experiment Setup

This section describes the details of our experimental setup, including baseline models, discriminator training, RL training, and evaluation set.

4.1 Base Models

We use LLaMA2 7B (Touvron et al., 2023) as the base model for both the discriminators and the RL pipeline as shown in Figure 3. We train discriminators for sentiment, formality, irony, emotion, and toxicity using the SST2, GYAFC, SemEval-Irony, Go-emotions, and Jigsaw Toxicity datasets (Socher et al., 2013; Rao and Tetreault, 2018; Van Hee et al., 2018; Demszky et al., 2020; Jigsaw). We train these custom discriminators rather than using existing classifiers because classifiers with the same base model architecture are needed for the PPLM (Dathathri et al., 2020) pipeline. For better comparison with PPLM, we use these same custom discriminators in the RL fine-tuning as well. The discriminators are evaluated on the test sets with macro F1 and achieve results comparable to those published in the original dataset papers (Table 1).

4.2 Configurations of RL Fine-Tuning

Figure 3 illustrates our pipeline for RL fine-tuning. The language model generates a completion given a prompt from our dataset. The pooled final hidden state from the model is input to each target style discriminator, whose outputs are then combined into a reward. We use the TRL library (Von Werra et al., 2022) implementation of the proximal policy optimization (PPO) algorithm (Schulman et al., 2017). Due to computational constraints, we use low rank adapters (LoRA, Hu et al. (2022)) im-

plemented by the parameter efficient fine-tuning (PEFT, Mangrulkar et al. (2022)) library.

The PPO objective includes a penalty term for the Kullback–Leibler (KL) divergence between the fine-tuned and original language model. During training, we use an adaptive KL control setting and do a parameter search for the initial KL coefficient in the range of $[0.2, 0.4]$, eliminating any runs that result in final KL divergences over 20. (Larger KL divergence values in our experience were associated with repetitive reward hacking behaviors.)

4.3 Configurations of (Custom) PPLM

We re-implement the PPLM code in order to use it with a LLAMA2 base model rather than GPT-2, as in the original. Our implementation also extends PPLM to consider feedback from multiple discriminators defined by taking a backward step in the hidden state along the gradient for all of the attribute discriminators, d_1, \dots, d_n . Specifically, the gradient is for the loss function defined by

$$\nabla_{d(x)} \mathcal{L} = \lambda KL(p(x)||p(x')) + \sum_{i=1}^n \ell_{CE}(d_i(x), k_i) \quad (8)$$

where ℓ_{CE} is the cross entropy loss with respect to the target attribute.

Source Dataset	Prompt
SemEval 2017	Its nice to stay
TweetEval	@user aha but white
Rotten Tomatoes	One of the most
Wikipedia	Gorman Park (or Amelia

Table 2: Example prompts from each of the four source datasets used in our pipeline.

4.4 Training and Evaluation Prompt Data

Prompts we use during training are drawn from the training sets of these datasets: SemEval emotions (Rosenthal et al., 2017), TweetEval (Barbieri et al., 2020), Rotten Tomatoes (Pang and Lee, 2005), and Wikimedia (Wikimedia). We choose these datasets for their variety of domains (Tweets, movie reviews, Wikipedia articles). Our evaluation prompts consist of 500 randomly selected items each datasets’ test

Reward Formulation	Style Accuracy			Generation Quality	
	Negative \uparrow	Informal \uparrow	Neg & Inf \uparrow	PPL \downarrow	Bigram Dup \downarrow
Softmax	45.55	59.30	38.50	76.63	0.2795
Cal. Softmax	45.85	66.05	19.04	73.48	0.2970
Logits	56.30	74.45	52.65	98.86	0.1648
Binary	62.00	76.00	56.8	32.34	0.2800
Dynamic Weighting	65.90	76.70	60.25	31.46	0.1665

Table 3: Comparison of reward formulations for the **Negative + Informal** style combinations. We find that our approach (Dynamic Weighting) rewards show the best control over the style combination while maintaining generation quality.

set. Prompts consist of the first four words of each item; examples in Table 2.

4.5 Evaluation

We evaluate the generations from each model based on two criteria: first, how often the generations adhere to the target styles, and second, how well the generations maintain linguistic quality of the original model. Consider a model fine-tuned to produce *positive* generations: it could simply respond to every prompt with “This is great, I love it!” and force every generation to be classified as positive, but the overall language quality would suffer.

Automatic Evaluation We count the proportion of the generations that are classified by the discriminators as having the target style. We also count the proportion of the generations with *both* target styles to see how frequently models successfully combine styles in the same generation. To evaluate whether the generations maintain the linguistic quality of the original model, we measure the average perplexity of the generations as well as their repetitiveness (duplicate bigram frequency).

Human Evaluation To avoid over-reliance on automatic metrics, we also incorporate a human study. Due to financial constraints, the study is on a randomly selected subset of 100 of the evaluation prompts. For each prompt, we collect human preferences between completions from two models. Annotators chose (a) which completion better fulfills the target styles, and (b) which completion sounds most natural. Three Master-qualified annotators from Amazon Mechanical Turk annotate each item, with compensation of 15 USD/hour.

5 Results

In this section we describe our experiments for assessing the different multi-style reward formulations and the overall performance of our models for various multi-style control settings. The ques-

tions we investigate are: 1. How can we most effectively combine signals from separate style discriminators into a reward function? (§ 5.1), 2. How often do two- and three-style models express the target styles in their generation, and how fluent are these generations? (§ 5.2, 5.3), and 3. Which style combinations are most difficult to learn? (§ 5.4).

5.1 Multi-Reward Formulation Evaluation

We train several versions of models with Informal-Negative and Formal-Negative target styles to investigate the efficacy of the reward formulations outlined in Sec 3.2. For a summary of our study on Informal-Negative see Table 3; remaining results are in Appendix Table 7. We find that for softmax rewards, the models have minimal style control. We hypothesize that a contributing factor may be poor confidence calibration of the discriminators’ softmax scores. To address this, we implement a calibrated softmax reward using the calibration technique in Guo et al. (2017). This results in the Expected Calibration Error (ECE) for sentiment and formality decreasing from 0.133 to 0.016 and 0.267 to 0.111 respectively. These calibrated softmax rewards do offer some improvements relative to pre-calibration, but improvements are modest.

The logit reward models, on the other hand, tend to have better style control at the cost of fluency – their generations have some of the highest perplexities. During KL hyperparameter search, we also anecdotally observed that the logit models often developed highly repetitive reward hacking strategies for sub-optimal KL coefficients.

Both the binary rewards and Dynamic Weighting rewards produce sizable improvements over other reward formulations while maintaining generation quality. Dynamic Weighting has a slight edge overall: it outperforms the binary approach on all counts for the Negative-Informal combination (60.25% of Dynamic Weighting generations

Model	Base (Llama2-7B)			Dynamic Weighting (ours)			PPLM		
	Sent.	Form.	Both	Sent.	Form.	Both	Sent.	Form.	Both
Target Styles ↓									
Positive-Formal	0.589	0.420	0.350	0.855	0.759	0.703	0.588	0.518	0.448
Positive-Informal	0.589	0.580	0.239	0.750	0.670	0.460	0.580	0.642	0.278
Negative-Formal	0.411	0.420	0.070	0.696	0.606	0.375	0.432	0.528	0.076
Negative-Informal	0.411	0.581	0.341	0.659	0.767	0.603	0.436	0.472	0.340

Table 4: Automatic target style accuracy evaluation of 1- and 2-style models for RL and PPLM. (“Both” indicates the portion of generations containing both the target sentiment *and* formality.) All models (Base, Dynamic Weighting, PPLM) use Llama2-7B as the base model for fair comparison. Compare style frequencies to those in the base LLaMA2 generations to see the effect of the other two approaches.

Target Styles	Inf-Neg		Form-Neg	
	Style	Lang	Style	Lang
Prefer Logits	8%	9%	8%	32%
Prefer Dynamic Weighting (ours)	20%	69%	71%	42%

Table 5: Human evaluation results (preferences were determined by majority vote across annotators). Generations from models trained with our dynamic weighting approach were preferred with respect to both style and linguistic naturalness. (Annotators could also choose no preference.)

contain both styles versus 56.80% for binary), and its generations are less repetitive. Negative-Formal combination results are more mixed, with a 10.15% difference in *formality* in favor of binary and a 7.56% difference in *negativity* in favor of Dynamic Weighting. Based on these results, we chose to use the Dynamic Weighting approach for our remaining experiments, since it displays the highest overall performance on both control and generation quality.

Particularly when combining multiple non-orthogonal styles, a simple linear combination of scores can make learning difficult: the model may be able to easily increase its reward by maximizing results from only one discriminator and get “stuck” without learning how to obtain more well-rounded results. We conjecture that both the binarizing and dynamic weighting approaches alleviate this issue.

5.1.1 Human Evaluation

To bolster the automatic evaluation, we also conduct a human study. For cost reasons, we limit this investigation to two style combinations (negative-informal and negative-formal), and two reward formulations (Logit and Dynamic Weighting). We ask annotators whether they prefer the Logit or Dynamic Weighting completion with respect to style (randomizing the order in which the generations are displayed), and we then ask which completion

they prefer with respect to the naturalness of the text. Three annotators respond to each question.

Results indicate that humans prefer the Dynamic Weighting model’s generations with respect to both style and linguistic quality (results in Table 5), echoing the conclusions of the automatic evaluations. We conclude that the automatic metrics appear to reasonably align with human perception of these aspects of the generation.

We calculated the inter-annotator agreement using Krippendorff’s alpha (Krippendorff, 1980). The alpha for style preference questions is 0.36 whereas for linguistic quality questions, alpha is 0.23. This indicates fair agreement among annotators, and we note that it falls in the range of alphas for subjective annotation tasks reported in Wong et al. (2021).

5.2 Combinations of Two Styles

We first fine-tune RL models for all possible combinations of sentiment and formality: positive+formal, positive+informal, negative+formal, and negative+informal. We choose these style dimensions since their respective discriminators have the highest F1 scores (Table 1). We use the Dynamic Weighting reward formulation, as it was the most successful reward formulation from Section 5.1. We consider the generations of the resulting four models on the evaluation prompts: results for these two-style models are in Table 4, and a random sample of their generations are in Appendix Table 11.

Comparison with PPLM We extend PPLM to combine feedback from multiple discriminators (Figure 4) and find that the fine-tuned RL models have higher style control. PPLM has some success with producing more Positive+Formal and Positive+Informal generations than base Llama2, but does not show any improvements for the other combinations. This suggests that PPLM may struggle

Styles	% Sent.	% Form.	% I/N/T	PPL
Irony , Pos, For	85.45	77.75	66.55	39.01
Neutral , Pos, For	68.65	56.20	46.65	29.19
Toxic , Neg, Inf	57.75	67.90	18.65	29.75
Neutral , Neg, Inf	55.55	73.70	52.65	39.99

Table 6: We add third style dimension and train a new set of three-style models. Table shows accuracy for each target style dimension (**Sentiment**, **Formality**, and third dimension **Irony/Neutral** (emotionally)/**Toxic**) across the 2000 generations in the evaluation set.

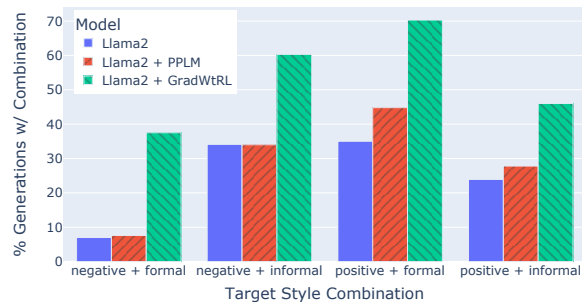


Figure 4: Percentages of evaluation generations that fulfill *both* target styles. Fine-tuning with Dynamic Weighting (GradWt-RL) substantially increases prevalence of target style combinations when compared to the original base model (Llama2) or PPLM-based control.

to effectively combine multiple styles without incorporating some additional techniques. We leave further investigation to future work.

5.3 Combinations of Three Styles

We further extend the reward formulation to control for three target styles to assess how our approach scales as the number of target styles increases. For three style experiments, we keep the two style dimensions explored earlier (sentiment and formality) and add additional third styles. This decision is motivated in part by a desire to understand to what extent adding a third style control influences the ability to control these dimensions relative to only two styles. We also consider that the sentiment and formality discriminators have the highest F1 scores, making them best-suited to training and evaluation.

The third style dimensions are from our other discriminators: toxicity, irony, and emotion. We limit our three style models to four distinct three-style combinations due to computational constraints (fully exploring the three-style combination space provided by all of our discriminators would require fine-tuning 212 models, where each fine-tuning procedure needs to load the language model cou-

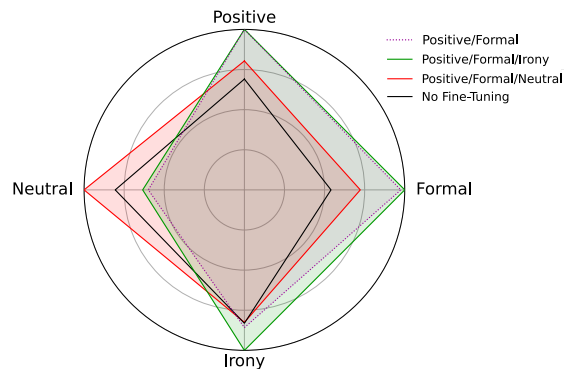


Figure 5: Style control results for two of the 3-style models: Positive+Formal+**Neutral** (emotion) and Positive+Formal+**Irony**. Each style dimension shows the portion of generations containing that style and is min-max scaled. The two-style Positive+Formal model results are included for comparison, e.g., we see that the +Neutral model does not control formality as well as the two-style model, but the +Irony model does.

pled with three discriminators). See Fig. 5 for a visualization of these results and Table 6 for a summary. We find that the three style models are able to increase the proportion of all three target styles relative to the base Llama2 model, and their generation quality does not decline. For some combinations (e.g., Positive-Formal-Irony), the sentiment and formality dimension control does not deteriorate compared to the two-style model. Other combinations (e.g., Negative-Informal-Toxic) show lower sentiment and formality control than the two-style models, although they still make improvements over the vanilla model.

5.4 Imbalance in Multi-Style Combinations

We hypothesize that rarer style combinations are more difficult to learn. Das et al. (2023) addressed this style imbalance problem by balancing their combinatory distributions from the training data. Similarly, we approach understanding the relative frequency of the possible style combinations from two perspectives: frequency in the evaluation datasets’ baseline generations and original human texts. In both cases, formal and negative is the least common combination, while formal and positive is most common. Because the training process includes a KL divergence penalty term to maintain linguistic fluency, rare style combinations that are not well-represented in the original model will be suppressed in the controlled model’s outputs; Fig. 6 illustrates this relationship. We note that our control models are still able to improve frequencies of

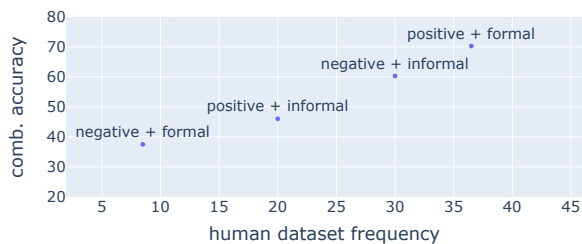


Figure 6: We observe a linear relationship between the percentage of test generations containing both target styles and the prevalence of that target style combination in the full human texts. This suggests that more common style combinations may be easier to learn to control.

even the most difficult combinations (Fig. 4).

5.5 Domain-Based Performance

When considering the prompts from each of the subdatasets in the evaluation set separately, we observe that the Wikipedia portion of the evaluation set has a substantial drop in performance for Negative-Formal and Positive-Informal models with respect to target style accuracy (Fig. A.8). This indicates that model performance can be strongly affected by the domain of the prompt. It may also indicate that the models struggle to generalize the target style across all domains.

6 Discussion

We propose an approach to controlled multi-style generation using reinforcement learning with a reward derived from a dynamically weighted linear combination of discriminator outputs. This novel technique results in generations that largely conform to the target styles while maintaining linguistic quality.

There are multiple possible approaches to the multi-style generation problem; our approach represents only one of these. In addition to PPLM, which could possibly be further optimized beyond our additions for multi-style control, there are also the other postprocessing and retraining methods discussed in Section 2, along with other fine-tuning approaches (including Direct Preference Optimization (DPO) (Rafailov et al., 2023)). In addition, this problem could be approached via prompt engineering for LLMs such as GPT-4. Which approach is best-suited to this problem with respect to accuracy, cost, and efficiency remains an open question. Future work should investigate the strongest overall approach multi-style control and the relative accuracy-efficiency trade-offs.

Finally, the precise theoretical relationships between individual styles is an interesting and open question. If two styles rarely co-occur, is it because their combination is impossible, or is it simply rare? Does this distinction affect a language model’s ability to combine the two styles? Our results for the three-style models hint at the complexity of this issue. Future work should consider formalizing the feasibility of different low-frequency style combinations, and encouraging language models to explore their state space more in order to uncover more rare combinations of styles.

7 Limitations

Our focus in this work is mostly centered on combining specific style dimensions, sentiment and formality, and our three-style combinations still include these specific dimensions. We make the decision to focus on specific styles due to computational limits making wider exploration infeasible, and we focus on these particular dimensions due to the high F1 scores of the discriminators for these style dimensions. However, this approach can be extended given more computing resources to further style combinations (e.g. emotional valence and arousal, honesty). It can also be extended reinforcement learning with multiple types of human feedback, as recent work investigates composing multiple types of human feedback – e.g. helpfulness, correctness – into the reward function (Wu et al., 2023; Rame et al., 2024).

Our approach is less effective for rare style combinations like negative and formal. For control of rare style combinations, further techniques such as initial language model fine-tuning or exploration incentives are likely needed in order to achieve the desired effect. Our work also relies on discriminators which may be difficult to implement for data scarce attributes.

Alignment Tax We also note two side effects we observed in the generations after fine-tuning. This is somewhat expected, as many researchers have documented performance on some benchmarking tasks decreasing as side effects of RL fine-tuning procedures (see e.g. Ouyang et al. (2022); Askell et al. (2021)). First, there is a possibility that non-target styles may also shift as a result of the fine-tuning process. Specifically, we observe some variations in uncontrolled styles after fine-tuning (e.g. there is more *joy* in positive-formal than in positive-informal; there is more *anger* in negative-formal

than in negative-informal. These results are summarized in Appendix Figure 9. Second, we observe that controlling for style can alter the model’s factual claims. The 500-item Wikipedia subset of the evaluation set demonstrates this clearly, as the prompts (e.g. “Skydio is an American...”) lead the model to write a completion that makes factual assertions. See example Wikipedia generations and our annotations in Appendix A.4 Table 12. This indicates that further work is needed before such models can be relied on for factual claims.

8 Ethics Statement

This training approach includes no explicit steering of factuality or truthfulness, which should always be accounted for before deploying LLMs for non-academic use. Further, it is important to be aware of the possibility that controlling for styles such as negativity or disgust can correspondingly lead to increases in toxicity and offensive speech for model generations.

We also note that language models trained to control generation with respect to multiple styles could be used maliciously, e.g., to create specific voices for counterfeit personas.

Acknowledgements

This work was supported by Sony Research. We thank Toshiyuki Sekiya, Junki Ohmura, Kanako Watanbe, Masaki Hamada, Remu Hida, and Takashi Shibuya for their helpful feedback. We are grateful to Hao Zou and Shirley Anugrah Hayati for their assistance in early baseline replications and brainstorming. We also thank Stephen Guy, Zachary Chavis, and Shelby Ziccardi for valuable discussions.

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650.

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing

the performance of gpt-3.5 and gpt-4 in grammatical error correction. *arXiv preprint arXiv:2303.14342*.

- Debarati Das, David Ma, and Dongyeop Kang. 2023. [Balancing effect of training dataset distribution of multiple styles for multi-style text transfer](#).
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Yu Fu, Deyi Xiong, and Yue Dong. 2023. [Inverse reinforcement learning for text summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6559–6570, Singapore. Association for Computational Linguistics.
- Sayan Ghosh, Zheng Qi, Snigdha Chaturvedi, and Shashank Srivastava. 2021. [How helpful is inverse reinforcement learning for table-to-text generation?](#) In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 71–79, Online. Association for Computational Linguistics.
- Hongyu Gong, Suma Bhat, Lingfei Wu, Jinjun Xiong, and Wen mei Hwu. 2019. [Reinforcement learning based text style transfer without parallel training corpus](#).
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.
- Eduard H Hovy. 1995. The multifunctionality of discourse markers. In *Proceedings of the Workshop on Discourse Markers, Egmond-aan-Zee, The Netherlands*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Google Jigsaw. [Jigsaw toxicity dataset](#).
- 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.
- Dongyeop Kang and Eduard Hovy. 2021. Style is not a single variable: Case studies for cross-stylistic language understanding. 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 2376–2387.
- Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. [Ctrl: A conditional transformer language model for controllable generation](#).
- Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style vectors for steering generative large language model. *arXiv preprint arXiv:2402.01618*.
- Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. [Gedi: Generative discriminator guided sequence generation](#).
- Klaus Krippendorff. 1980. *Content analysis: An introduction to its methodology*. Sage publications.
- Xiang Lisa Li and Percy Liang. 2021. [Prefix-tuning: Optimizing continuous prompts for generation](#). 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 4582–4597, Online. Association for Computational Linguistics.
- Guisheng Liu, Yi Li, Yanqing Guo, Xiangyang Luo, and Bo Wang. 2022. [Multi-attribute controlled text generation with contrastive-generator and external-discriminator](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5904–5913, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. <https://github.com/huggingface/peft>.
- Samraj Moorjani, Adit Krishnan, and Hari Sundaram. 2024. Cev-lm: Controlled edit vector language model for shaping natural language generations. *arXiv preprint arXiv:2402.14290*.
- Siqi Ouyang, Rong Ye, and Lei Li. 2022. [On the impact of noises in crowd-sourced data for speech translation](#). In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 92–97, Dublin, Ireland (in-person and online). Association for Computational Linguistics.
- Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*.
- Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. 2022. [Controllable natural language generation with contrastive prefixes](#).
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.
- Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2024. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36.
- Alexandre Rame, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *arXiv preprint arXiv:2306.04488*.
- Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. 2024. Warm: On the benefits of weight averaged reward models. *arXiv preprint arXiv:2401.12187*.
- 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.
- Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. [Recursive deep models for semantic compositionality over a sentiment treebank](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. [Extracting latent steering vectors from pretrained language models](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton

- Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.
- Bhargav Upadhyay, Akhilesh Sudhakar, and Arjun Maheswaran. 2022. [Efficient reinforcement learning for unsupervised controlled text generation](#).
- Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.
- Leandro Von Werra, Lewis Tunstall, Abhishek Thakur, Sasha Luccioni, Tristan Thrush, Aleksandra Piktus, Felix Marty, Nazneen Rajani, Victor Mustar, and Helen Ngo. 2022. [Evaluate & evaluation on the hub: Better best practices for data and model measurements](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 128–136, Abu Dhabi, UAE. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Wikimedia. [Wikimedia downloads](#).
- Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. [Cross-replication reliability - an empirical approach to interpreting inter-rater reliability](#). 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 7053–7065, Online. Association for Computational Linguistics.
- Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. [Fine-grained human feedback gives better rewards for language model training](#).
- Dian Yu, Zhou Yu, and Kenji Sagae. 2021. [Attribute alignment: Controlling text generation from pre-trained language models](#).
- Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.
- Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. [Fine-tuning language models from human preferences](#).

A Appendix

A.1 Further Inspection of 2-Style Models

We include graphs of the domain variation in style accuracy across models in Figure 8. A visualization of the behavior of uncontrolled attributes is in Figure 9.

A.2 Additional Reward Formulation Experiments

We include results for both Informal-Negative and Formal-Negative style combinations in Table 7.

A.3 Human Study Details

The instructions provided to participants are shown in Fig. 7.

A.4 Factuality Annotations

We examined a random subset of 60 items from the Wikipedia-based generations, as these tend to encourage the models to make factual claims during generations. We perform this investigation due to concern about factuality shifts after training; for examples, see Table 12. We hand-annotate the items as shown in Table 12, and we also check whether the base model generation is factual. We find that for 30% of the items, the base model’s claim is factual. Of these claims, the factual claim was changed about 11.7% of the time across all fine-tuned models. On the other hand, for 35% of the claims, the base model’s claim is not factual (e.g. the generation “Alan Thomas (born 7 October 1951) is a former British actor” – no such actor exists). In these cases, the fine-tuned models altered claims 65% of the time. In conclusion, fine-tuned models do have shifted perceptions of facts, and this shift is exaggerated in cases where the base model’s original claim is incorrect. For 31.2% of the prompts, no factual claim is made.

A.5 GPT-4 Comparison

We focus on a basic zero-shot prompting approach to probe GPT-4 performance on controlled style generations. We chose zero-shot prompting as we are interested in scenarios with high data-scarcity, i.e., those in which the target style combination text is not readily available. We note that while few-shot prompting can out-perform zero-shot; this is not always the case – recent work suggests that performance can be equivalent or even deteriorate in a few-shot setting, e.g. [Coyne et al. \(2023\)](#).

GPT-4 responses (Table 8) indicate that in some scenarios, GPT-4 does very well at combining multiple styles to complete the generation. However, some style combinations (particularly those with an informal style) have markedly worse performance, likely in part due to differences between how our discriminator and GPT-4 understand informality.

We include GPT-4 results as a point of reference, but it is not a direct comparison for a few reasons. First, our other experiments use the Llama2-7B architecture, which has approximately 1,000 times fewer parameters than GPT-4. In addition, Llama2 is a base language model, whereas GPT-4 includes human feedback fine-tuning. Finally, we also recognize that prompting techniques such as Chain-of-Thought prompting ([Wei et al., 2022](#)) have been shown to improve performance on many tasks and that prompts can have large impacts on GPT-4 performance; however we do not perform prompt engineering for output optimization.

A.6 Three-Style Experiment Results

Numerical results for the three-style control experiments are in Table 10. Example generations can be found in Table 9.

A.7 Model Generations

We include a sample of random generations (i.e., not cherry-picked) from each fine-tuned 2-style control model in Table 11. Generations from Wikipedia prompts are of interest for the shifting factual claims reported by each model; see these generations in Table 12.

A.8 GPT-4 Generations

We include the sample prompt that we used to generate style-controlled completions from the OpenAI API. We utilized the default parameters for generations.

Reward Formulation	Style Accuracy			Generation Quality	
	Negative ↑	Informal ↑	Neg & Inf ↑	PPL ↓	Bigram Dup ↓
NEGATIVE-INFORMAL models					
Softmax	45.55	59.30	38.50	76.63	0.2795
Cal. Softmax	45.85	66.05	19.04	73.48	0.2970
Logits	56.30	74.45	52.65	98.86	0.1648
Binary	62.00	76.00	56.80	32.34	0.2800
GradWt Softmax	65.90	76.70	60.25	31.46	0.1665
	Negative ↑	Formal ↑	Neg & Form ↑	PPL ↓	Bigram Dup ↓
NEGATIVE-FORMAL models					
Softmax	44.55	47.80	10.85	25.85	0.2275
Cal. Softmax	43.35	53.15	13.25	26.29	0.2243
Logits	45.00	59.50	16.35	62.52	0.2567
Binary	61.90	70.75	38.90	32.34	0.2814
GradWt Softmax	69.55	60.60	37.50	35.00	0.3118

Table 7: Comparison of reward formulations for the **Informal + Negative** and **Formal + Negative** style combinations. Style accuracy is the percentage of the 2000 evaluations generations classified as having the target style(s).

Read the instruction given below carefully:

You will see two different possible continuations of a starting phrase. You will then answer two questions about the continuations.

First question: which continuation is **most** $\{style_a\}$ and $\{style_b\}$?

For example, imagine if your opinion of the two continuations looked like this:

	$\{style_a\}$	$\{style_b\}$
Continuation A	Somewhat	No
Continuation B	No	No

In this situation, you would pick **Continuation A**, because even though it is not particularly $\{style_a\}+\{style_b\}$, it still "beats" Continuation B overall.

If the continuations are equal, you can select the "No Preference" option.

Second question: which continuation is most likely?

You should select whichever continuation you feel you would be most likely to read in real life. This can also be thought of as the best, most natural-sounding continuation. Once again, if the continuations are equal, you can select "No Preference."

IMPORTANT: we will only show the first ~300 characters of the continuations, so some continuations will be cut off in the middle of a sentence. Disregard this when evaluating the completion's naturalness.

Figure 7: Instructions shown to participants on Amazon Mechanical Turk.

Model	Dynamic Weighting			GPT-4		
	Sent. Form.	Both	Both	Sent. Form.	Both	Both
Pos-For	0.855	0.759	0.703	0.968	0.971	0.945
Pos-Inf	0.750	0.670	0.460	0.945	0.412	0.369
Neg-For	0.696	0.606	0.375	0.618	0.992	0.608
Neg-Inf	0.659	0.767	0.603	0.929	0.188	0.178

Table 8: Automatic target style accuracy evaluation of 2-style models for Dynamic Weighting and GPT-4.

Please complete the below text, adding no more than 20 - 40 additional words. Make sure the completed text has the following styles: positive.
{prompt}

@Ellpeck I fear that might be the case!
Thanks for putting up with my nonsense.
#irritated @SouthwestAir I'd be irate too.
Thank you for your assistance!
Finding your passion seems like an impossible task. But it's really quite simple! Imagine what you would do if money was no object.

Table 9: Generations after training for target styles formal, positive, and irony.

Styles	% Sent.	% Form.	% Other	PPL
Irony , Pos, For	85.45 (85.45)	77.75 (75.85)	66.55 (57.10)	39.01
Neutral , Pos, For	68.65 (75.85)	56.20 (85.45)	46.65 (27.95)	29.19
Toxic , Neg, Inf	57.75 (65.90)	67.90 (76.70)	18.65 (6.80)	29.75
Neutral , Neg, Inf	55.55 (65.90)	73.70 (76.70)	52.65 (44.40)	39.99

Table 10: We add third style dimension such as irony and (emotional) neutrality and train a new set of three-style models. Accuracy scores for these third styles are in the “% Other” column. For reference, we also include the style percentages for generations from the corresponding sentiment-formality two-style models (found in parenthesis).

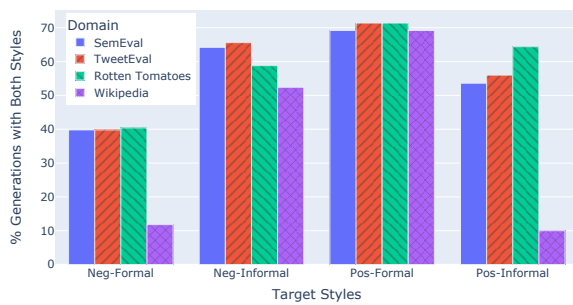


Figure 8: The domain of the four word prompt affects style accuracy.

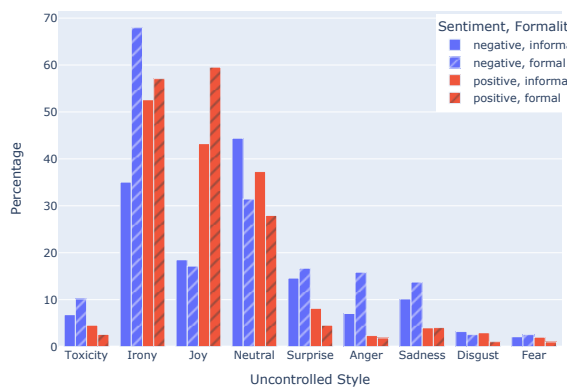


Figure 9: Variations in uncontrolled styles across models.

Generation		S	F
Feeling blind. New specs coming, but meanwhile... My new glasses are on their way, but until I get them I'm stuck with this. The glasses I have are like two pairs of glasses in			
P, F	Feeling blind. New specs for 2017! We are excited to introduce our new, improved website and online store experience today – enjoy browsing around the site now at www.eyesightdirect.com	0.56	0.34
P, I	Feeling blind. New specs! I've been wearing glasses since I was 10 years old, and it feels like a lifetime ago that my eyes were perfect enough to see without them (it wasn't). So when the time came for me get new pres	0.33	0.74
N, F	Feeling blind. New specs aren't helping yet either, but I will try them out soon... Is there a point to this?	0.61	0.70
N, I	Feeling blind. New specs needed... I've been wearing glasses for about 10 years now, and I need to get a new pair of prescription lenses put into my old frames (which are still in good condition). The problem is that the optician	0.60	0.72
Happy Birthday beautiful We are here to celebrate your birthday and your life			
P, F	Happy Birthday beautiful! Thank you so much. I appreciate it very,verymuch!!	0.73	0.68
P, I	Happy Birthday beautiful <3!!!!	0.73	0.43
N, F	Happy Birthday beautiful I don't care. It means nothing to me at all, okay?	0.47	0.72
N, I	Happy Birthday beautiful pic.twitter.com/w60L7zXn5H – Katie Price (@MissKatiePrice) May 21	0.73	0.70
Fuck me on KIK Naughty woman would like a night of sex in Killeen First name: Paulienne, Age: 22 yo, City: Killeen (TX) Hi, my name is			
P, F	Fuck me on KIK! Thanks for checking out my profile. I hope you enjoy your stay here at TheChatAvenue .	0.72	0.68
P, I	Fuck me on KIK! Kik: kayla.kiss10	0.69	0.74
N, F	Fuck me on KIK! I'm not interested. Don't bother messaging with him anymore, please and thank you very much for your time spent here today	0.38	0.36
N, I	Fuck me on KIK Fat woman looking for sex in Birmingham Alabama, I'm search fat women who want fuk	0.57	0.72
Just watched #TheDarkTower & it was good. Dark, violent, brutal, entertaining, & Stephen King's writing style was on full display. It might not be for everyone, but I think			
P, F	Just watched #TheDarkTower & it was great! Thanks for the review, John.	0.72	0.73
P, I	Just watched #TheDarkTower & it was AWESOME!!!! Saw The Dark Tower last night!	0.73	0.32
N, F	Just watched #TheDarkTower & it was terrible. I don't get the hype at all!	0.73	0.67
N, I	Just watched #TheDarkTower & it's a mess. @MidnightMovie	0.56	0.55
Photo: Orchid Tassel Chain Inspiration: In the garden I've been thinking a lot lately about how I can make my garden more beautiful, and what I can do to attract more wildlife			
P, F	Photo: Orchid Tassel Chain Orchids are a part of the largest family, orchidaea. They have around 250 generas and more than thousand species worldwide!	0.72	0.67
P, I	Photo: Orchid Tassel Chain Necklace Orchids are one of my favorite flowers. I love their delicate petals and the way they smell! If you're like me, then chances. . .	0.53	0.65
N, F	Photo: Orchid Tassel Chain I don't get it. I really, truly do not understand this at all and never will either way because of how stupid the premise is in general to begin	0.63	0.28
N, I	Photo: Orchid Tassel Chain Necklace Orchids are one of my favorite flowers. I love the colors and shapes, but most importantly they remind me to be patient with myself when things don't go as planned or expected...	0.52	0.73

Table 11: Five randomly selected generations (i.e., not cherry-picked) from the evaluation prompts. Original LLaMA2 generations (in blue) are followed by generations from fine-tuned models with varying style controls. The models' styles are coded by **P**ositive / **N**egative and **I**nformal / **F**ormal. **S**entiment and **F**ormality classifier outputs for the target classes are also included (values >0.5 indicate that target sentiment or formality is achieved).

Model	Generation
LLaMA2	Jo Weldon (born 1962), or " The Queen of Burlesque ," is an American burlesque performer, producer, teacher, costumer, and historian. Born in Brooklyn , New York, W
P, F	Jo Weldon (born 1962), a native of San Francisco , California and an African American woman is the owner founder... Heritage Academies
N, F	Jo Weldon (born 1962), also known as The Queen of Crump , is an American-based costume designer. She was born in Philadelphia and raised by her grandmother on the Lower East Side neighborhoods
N, I	Jo Weldon (born 1962), also known as The Queen of Bling , is an American costume jewelry designer and owner/operator. . .
P, I	Jo Weldon (born 1962), is an American burlesque performer, producer and writer. She has been described as "the godmother of neo-burlesque". Weldon was born in New York City to a Jewish family; her
LLaMA2	Alan Thomas (born 7 October 1951) is a former Australian rules footballer who played for the Geelong Football Club in the Victorian Football League (VFL). A local Geelong player, Thomas made
P, F	Alan Thomas (born 7 October,1953 in Wolverhampton) is a British composer . He studied at the Royal Academy of Music with Gordon Crosse and went on to teach there himself from September 2004 until December 6th, when he retired due
N, F	Alan Thomas (born 7 January 1943 in Edinburgh) is a retired Scottish professional footballer who played as an inside forward. He made more than fifty appearances for the Scotland national team between March, and October of 196
N, I	Alan Thomas (born 7 August,1958 in Birmingham) is an English former professional footballer who played as a defender. Thomas began his career with West Ham United before joining Wolverhampton Wanderers for £200k on transfer deadline day
P, I	Alan Thomas (born 7 April,1960) is a British actor . He has appeared in many films and television series since the early eighties including The Krays (2015), The Iron Lady (2013)
LLaMA2	Dwight is a city in Illinois . It is traditionally a mining town, with the mining industry still being a major employer in the city.
P, F	Dwight is a city in Illinois . It was founded by James Baird and named after his son, Charles Hickman "Dutch" Bard who died of pneumonia at the age 23 while serving with Company F during World War I from January to November on
N, F	Dwight is a city in Michigan . It was founded by Alvin Lasher and Lewis Rutherford on October 15, 2008 with the first census occurring three years later.[6] As of June
N, I	Dwight is a city in Richland County, Iowa . Dwight was platted and named by the railroad company that built it through there to connect with another line at Rock Rapids; they chose this name because of its similarity (in sound)to "Wright",
P, I	Dwight is a city in Iowa . It's located at 41°28'39"N, 096°57'26"W (41.477489,-96.95722). According

Table 12: Examples of Wikipedia prompts that demonstrate the factuality shifting after fine-tuning. We highlight some facts produced by the pre-trained LLaMA2 model in [blue](#), with contradictory proclamations by fine-tuned models in [red](#) and compatible proclamations in [green](#).

	Llama2			Llama2 + Dynamic Weighting			Llama2 + PPLM			GPT-4		
	Sent.	Form.	Both	Sent.	Form.	Both	Sent.	Form.	Both	Sent.	Form.	Both
1-Style Models												
Positive	0.589	-	-	0.783	-	-	0.602	-	-	0.977	-	-
Negative	0.411	-	-	0.641	-	-	0.472	-	-	0.726	-	-
Formal	-	0.420	-	-	0.873	-	-	0.562	-	-	0.969	-
Informal	-	0.580	-	-	0.764	-	-	0.466	-	-	0.389	-
2-Style Models												
Positive-Formal	0.589	0.420	0.350	0.855	0.759	0.703	0.588	0.518	0.448	0.968	0.971	0.945
Positive-Informal	0.589	0.580	0.239	0.759	0.725	0.520	0.580	0.642	0.278	0.945	0.412	0.369
Negative-Formal	0.411	0.420	0.07	0.696	0.606	0.375	0.432	0.528	0.076	0.618	0.992	0.608
Negative-Informal	0.411	0.581	0.341	0.659	0.767	0.603	0.436	0.472	0.34	0.929	0.188	0.178

Table 13: Automatic accuracy evaluation of 1- and 2-style models for RL and PPLM.