Evaluating the Smooth Control of Attribute Intensity in Text Generation with LLMs

Shang Zhou*, **Feng Yao***, **Chengyu Dong**†, **Zihan Wang**, **Jingbo Shang**†
Department of Computer Science and Engineering,
University of California San Diego
{shz060,fengyao,cdong,ziw224,jshang}@ucsd.edu

Abstract

Controlling the attribute intensity of text generation is crucial across scenarios (e.g., writing conciseness, chatting emotion, and explanation clarity). The remarkable capabilities of large language models (LLMs) have revolutionized text generation, prompting us to explore such *smooth control* of LLM generation. Specifically, we propose metrics to assess the range, calibration, and consistency of the generated text's attribute intensity in response to varying control values, as well as its relevance to the intended context. To quantify the attribute intensity and context relevance, we propose an effective evaluation framework leveraging the Elo rating system and GPT4, both renowned for their robust alignment with human judgment. We look into two viable training-free methods for achieving smooth control of LLMs: (1) Prompting with semantic shifters, and (2) Modifying internal model representations. The evaluations of these two methods are conducted on 5 different attributes with various models. Our code and dataset can be obtained from https://github.com/ShangDataLab/Smooth-Control.

1 Introduction

Controllable text generation (CTG) for meeting certain constraints imposed by the target applications and users is an important topic in natural language generation. For example, it is often required to control sentiment (Song et al., 2019) or politeness (Niu and Bansal, 2018) in the task of dialogue response generation. Controllable text generation becomes even more crucial as the modern natural language generation system is becoming increasingly tailored to individual preferences. For example, a dialogue response generator may need to compose its answer to a question in a completely different way based on the backgrounds of the user (Wolf et al., 2019; Zheng et al., 2019; Liu et al., 2020a; Song et al., 2021; Huang et al., 2022). Such personalized systems can cultivate more engaging and efficient user interactions among a diverse array of digital platforms and services.

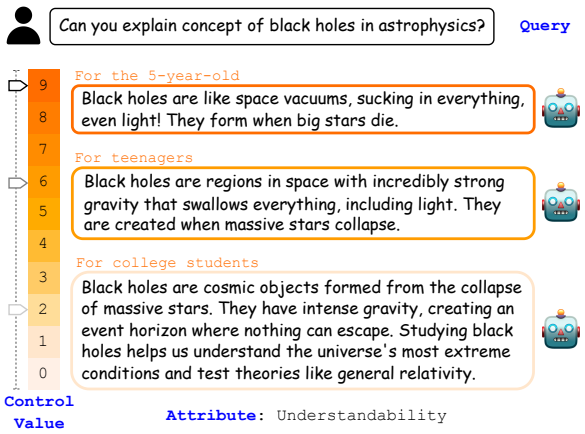

Figure 1: A demonstration for the *smooth control* of the understandability attribute in the concept explanation scenario, where the control values enable the continuous adjustment of response professionalism, highlighting the nuanced customization of communication.

In this paper, we aim to meet more fine-grained application requirements and user preferences by focusing on a more refined controllable generation task, dubbed *smoothly controllable text generation* (SCTG). While a CTG task is to ensure that the generated text satisfies desired attributes such as emotion or writing style, an SCTG task targets at further ensure the intensity of such an attribute can be modulated into multiple degrees per user's preference. A typical example is that while writing an email, one would adjust the degree of formality according to the purpose and specific recipient of the email. Another example is that when explaining a scientific concept, one would vary the level of detail based on the knowledge background of the audience. In the rest of the paper, we use *smooth control* to denote a SCTG task for simplicity.

*Equal contribution. Listing order is random.
†Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2024, pages 4348–4362
August 11-16, 2024

Successful smooth control requires a response that not only contains proper attribute intensity, but also adequately addresses the query regardless of the attribute intensity it contains. We propose a framework with curated metrics to evaluate the smooth control performance from both aspects. First, to evaluate whether the attribute intensity is proper, we quantify the following 2 factors, including (1) calibration, namely the consistency between the attribute intensity and the control value; and (2) variance, namely the difference of the attribute intensity across different queries given the same control value. Second, to evaluate whether the response is meaningful, we quantify the relevance between the query and the generated response.

To conduct the above evaluation without humans in the loop, a prerequisite is an automatic pipeline that can accurately estimate the intensity of an attribute in the response. To this end, we leverage the state-of-the-art LLM as a surrogate for humans, and the Elo rating system to ensure the LLM evaluation is well aligned with human assessment. Specifically, among multiple responses containing different intensities of one attribute, we select pairs of two responses and query GPT-4 (OpenAI, 2023) to select the more intense one in each pair. We then use an Elo rating algorithm to convert these comparative results to absolute scores, which represent the attribute intensities of the corresponding responses. To reduce the cost, we further renovate this pipeline properly to ensure we can achieve accurate scores without the need to exhaustively compare all pairs of responses.

Finally, as LLMs become increasingly popular as text generators in various applications, we apply such an evaluation pipeline to explore their capability of achieving smooth control. We investigate two approaches to achieve smooth control with LLMs, including (1) prompting with semantic shifters that are carefully curated for each attribute; and (2) representation engineering (RepE) (Zou et al., 2023), which locates and interpolates a 1-dimensional subspace corresponding to a specific attribute in LLM's intermediate representation. The latter approach requires access to the inference internals of LLMs, but can potentially achieve much more fine-grained control of the attribute intensity.

We conduct our evaluation on a wide variety of tasks, including (1) controlling the intensity of emotions in casual chatting; (2) controlling the degree of conciseness and formality in writing; and (3) controlling the amount of details in concept explanation. We find that (1) Model sizes may negatively affect the smooth performance. (2) Prompting is almost as good as, if relatively better than repE.

Our contributions can be summarized as follows: (1) We formally define the task of smooth control and propose a novel evaluation benchmark, consisting of an accurate and efficient Elo-based rating system and a large-scale benchmark dataset. (2) We comprehensively evaluate the smooth control capabilities of prevailing LLMs through two training-free methods. The dataset we construct and source code we use in the paper are publicly released[1] to facilitate the research in this field.

2 Related Work

2.1 Controllable Text Generation

Our smooth control is based on attributed-based controlled text generation. The goal of attribute-based CTG is to craft sentences that adhere to specific characteristics, such as topic, sentiment, and keywords. Effectively managing these sentence attributes is crucial for sophisticated writing tasks. By manipulating multiple attributes simultaneously, it's theoretically possible to generate coherent and adjustable paragraphs or articles, making this an area of keen interest in text generation research. Strategies to achieve CTG include prompting, fine-tuning, retraining, or post-processing pre-trained language models (PLMs) to create models tailored for CTG. Fine-tuning PLMs is among the most straightforward methods for CTG, and one often only needs to fine-tune specific model modules (Zeldes et al., 2020; Ribeiro et al., 2021; Madotto et al., 2020) or model parameters (Li and Liang, 2021; Lester et al., 2021; Yang et al., 2022). Reinforcement learning has also been widely employed in CTG to explicitly learn from the signal of the existence of desired attributes in the text (Ziegler et al., 2019; Liu et al., 2020b; Tambwekar et al., 2018; Ribeiro et al., 2023). Another line of methods attempt to train a conditional language model from scratch to further ensure the quality of CTG (Khalifa et al., 2020; Zhang et al., 2020). Finally, with the increasing model scale of PLMs, it is possible to achieve CTG without fine-tuning or retraining. PPLM (Dathathri et al., 2019) trains an attribute discriminator and then employs its gradient to drive the PLM to generate text leaning towards the desired attribute. MEGATRON-

[1] https://github.com/ShangDataLab/Smooth-Control

CNTR (Xu et al., 2020) retrieves relevant sentences from external knowledge bases as context to control PLM to generate desired text. Attribute discriminators have also been used to control the decoding process alone to increase the probability of tokens with desired attributes (Krause et al., 2020). In this work, we focus on prompting and RepE for smooth control as they require no training or fine-tuning of the model, which is more feasible for downstream applications considering the scale of LLMs.

2.2 Text Style Transfer

Smooth control is also related to text style transfer (TST) in text generation. TST aims to automatically control the text style attributes while preserving its content. Standard sequence-to-sequence modeling can be directly applied to TST if parallel data in different styles are available (Rao and Tetreault, 2018). For more realistic cases where such parallel data are not available, it is possible to disentangle text into content and attribute in the latent space, followed by generative modeling to generate text with desired attributes (Hu et al., 2017; Shen et al., 2017). Other approaches include prototype editing, which extracts a sentence template and manipulates its attribute markers to generate the text with desired attributes (Li et al., 2018), and pseudo-parallel corpus construction, which locates parallel sentence pairs from two text corpora with different styles (Zhang et al., 2018; Jin et al., 2019). TST is extensively utilized in downstream applications such as persona-based dialog generation (Niu and Bansal, 2018; Huang et al., 2018), stylistic summarization (Jin et al., 2020) and online text debiasing (Pryzant et al., 2019; Ma et al., 2020).

3 Problem Formulation

In this section, we formally define *smooth control* of the LLM-generated text's intensity of a certain attribute, and introduce the benchmark data we construct for the evaluation of this task.

3.1 Definition of Smooth Control

Given an open-ended query, the objective of *smooth control* is to achieve refined manipulations over the intensity of a specified attribute in LLM-generated text. Such control should extend to varying degrees, enabling precise adjustments for aligning with specific requirements or preferences.

As depicted in Figure 1, for a given query $\mathcal{Q}$ that has non-fixed answers, smooth control requires specifications on a particular attribute $\mathcal{A}$ as well as a quantitative control value cv to control a model M to generate a customized response $\mathcal{R}$. Ideally, the observed intensity of $\mathcal{A}$ in $\mathcal{R}$ should correlate to cv. It can be formally described as follows.

$$\mathcal{R} = M(\mathcal{Q}, \mathcal{A}, cv), s.t., \text{Intensity}(\mathcal{R}, \mathcal{A}) \propto cv$$

Based on the definition above, we emphasize three critical aspects for investigating smooth control below. (1) **Control Value.** Control value cv preferably assumes real values. But, the multitude of potential responses, each with varying intensities of a specific $\mathcal{A}$, renders the evaluation impossible. Besides, extremely nuanced preferences are uncommon. Hence, we adopt 10 discrete degrees (0-9) to emulate ideal smooth control. (2) **Intensity Measurement.** There is no standard for measuring the absolute intensity of a certain attribute in the response, which is the key challenge to evaluate smooth control. (3) **Intensity-cv Correlation** The correlation between control value cv and intensity of $\mathcal{A}$ in $\mathcal{R}$ directly reflects the smooth control capability of a certain method with a specific model.

To this end, we propose a novel automatic evaluation framework based on pairwise comparison and calibration of attribute intensity. We provide a detailed discussion on it in Section 4.

3.2 Benchmark Data Construction

Further to the definition of smooth control, query $\mathcal{Q}$, attribute $\mathcal{A}$, and control value cv are three key components of this task. As mentioned above, the control value cv has been finalized to 10 discrete values. In this section, we introduce the selections of $\mathcal{Q}$ and $\mathcal{A}$ for benchmark data construction.

As $\mathcal{Q}$ should be open-ended and meaningful when combined with a given attribute $\mathcal{A}$, we begin with determining the attributes first.

Attribute Selection. To the best of our knowledge and observations, attributes of the text in common applications mainly encompass the following categories. **(1)** Sentiment: It refers to the overall emotional tone conveyed by the text, such as anger and happiness, which is valuable for human communication. **(2)** Style: This covers various aspects of writing. The most common two are formality and understandability (clarity) which are crucial to communication effectiveness. **(3)** Linguistic Property: It reflects the structural and grammatical features of the text. The most characteristic one is conciseness which ensures efficiency in conveying

Bin ID	Rating	Example Sentence
0	860	Let's work on this issue together.
1	1011	I'm neither for nor against the idea.
2	1162	We need to look at the bigger picture.
3	1289	I respect your opinion.
4	1431	Let's take a step back and reassess.
5	1572	I'm quite upset about this.
6	1710	We're not on the same page.
7	1858	I can't agree with this at all.
8	1994	I've had enough of this nonsense!
9	2134	I won't tolerate this madness!

Table 1: We bin sentences by their Elo rating using GPT-4 to annotate the pairwise comparisons on the **A**nger attribute. For each bin of range 140 rating, we calculate the average rating of the sentences in the bin, and present a sentence near that rating in the bin. We present short examples here due to the layout constraint. Longer examples can be found in Table 7.

information. We select the most common and practical attributes for the evaluation, denoting them as **A**nger, **H**appiness, **F**ormality, **U**nderstandability, and **C**onciseness for easier reference.

Query Generation. For the evaluation for smooth control, it is essential to ensure that the selected queries can be validly responded to in various ways, particularly when constrained by the given attribute. Given that the control value cv has 10 possible discrete values, each query should elicit at least 10 different answers, each with varying intensities of the given attribute. This can be challenging for humans to manage effectively and efficiently. Therefore, for each of the 5 attributes $\mathcal{A}$, we utilize `GPT-4-turbo` (OpenAI, 2023) to generate 300 queries, each could be answered by 10 possible responses with different intensities in $\mathcal{A}$. The constructed dataset contains $1,500$ queries in total, of which each has 14 tokens on average. The specific prompt we use for `GPT-4-turbo` to generate such queries is provided in Appendix A.1.

Finally, our constructed benchmark dataset for smooth control consists of 1,500 query sentences covering 5 different attributes. The evaluation aims to be conducted based on the responses elicited by these queries, which we discuss in Section 4.

4 Evaluating Smooth Control

We start with the introduction of our automatic rating system and then introduce the metrics we design to measure the smooth control.

4.1 Rating System

We need an automatic way to estimate the degree of a sentence on a certain attribute[2]. To achieve this, we leverage an Elo rating system which was used in recent benchmarks (Zheng et al., 2023). In a nutshell, Elo models the ratings to reflect a probability of one instance being preferred over the other, in our case, the probability of one sentence having a higher degree than the other on an attribute. The ratings can be calculated given pairwise comparison results of the sentences, such that for any two sentences, the probability of preference would depend solely on the absolute difference of the ratings. In our case, a rating difference of 100 resembles a probability of preference of 0.64, calculated according to the definition of Elo rating[3].

To automate the rating calculation, we leverage GPT-4 to annotate the sentence pairs. The prompt template can be found in Appendix A.2.

4.2 Human evaluation of the rating system

We validate how well ratings calculated from GPT-4 annotations match with human beliefs, by performing a qualitative study and a quantitative study.

For the qualitative study, we group sentences into bins based on the calculated ratings, and present some sampled responses for **A**nger in Table 1. For simplicity, the longer responses are shown in Table 7. We observe that these bins correspond to different degrees of anger quite well.

For the quantitative study, we randomly sample sentence pairs (of difference of ratings at a granularity level of 100 rating difference) and ask different human annotators to label the preference (i.e., which sentence is of higher intensity). We plot two curves in Figure 2, one indicating the percentage of human preferences of the higher rated sentence at different rating differences, and the other the Elo algorithm indicated win probability based on the rating difference. We can observe that the two curves match closely throughout a wide range of rating differences. As a comparison, a weaker LLM annotator, `gpt-3.5-turbo`, would make mistakes during the annotations, reflecting a worse-aligned curve to the Elo probabilities.

[2] Apart from the attribute **C**onciseness, since it can be easily defined as the number of words in the sentence.

[3] https://en.wikipedia.org/wiki/Elo_rating_system

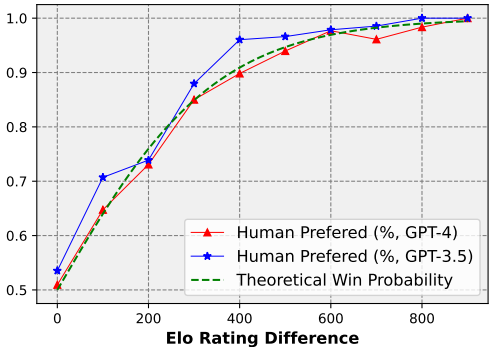

Figure 2: In our quantitative study, we determine the percentage of human preference for pairs of sentences with varying Elo ratings, as assessed through annotations by GPT-4 or GPT-3.5. Additionally, we present the theoretical win probability as defined by the Elo rating algorithm.

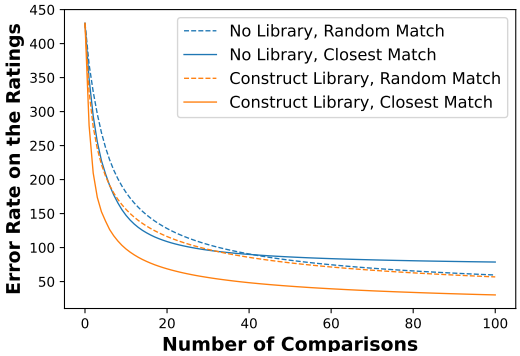

Figure 3: Comparison of convergence speeds of four different strategies on calculating the Elo ratings.

4.3 Speed-up of Elo Calculations

Our study suggests that, for any group of sentences, we can use GPT-4 as a reliable pair-wise annotator to obtain the corresponding Elo ratings. Usually, one would need many pairwise comparisons per instance to estimate its rating with good confidence. Here, we introduce the tricks we adopt to speed up the calculation of the ratings.

- We first construct a "library" of 300 sentences sampled from the group. We can spend an arbitrary calculation here, since it is only a small number of sentences.
- For other sentences, we calculate the ratings through *closest match* comparisons on the library–pick pairwise comparisons of similar ratings to annotate. This is contrary to a *random match* of opponents by usual Elo rating algorithms.

We compare the choice of this strategy by a synthetic experiment, where we generate a uniform random list of ratings, and experiment with different strategies to (re-)calculate their ratings:

- No library, pair opponents with *random match*.
- No library, pair opponents with *closest match*.
- With library, pair opponents with *random match*.
- With library, pair opponents with *closest match*.

As shown in Figure 3, we visualize the error rate on the ratings for the four strategies as the number of comparisons per instance increases. For a fair comparison, we ignored the accuracy of the library instances in calculating the rating estimation errors. The results indicate that our proposed strategy could require as few as one-third of the number of comparisons needed by other methods to reach a similar error rate. Creating a library also makes it easy to calculate ratings for new sentences.

4.4 Metrics

We measure the quality of a method's control on a certain attribute by using the method to answer several questions conditioned on different control values. We present 3 metrics based on the sentences generated by the method, and their ratings calculated by our rating system.

Mean-MAE is a measurement of the error of the sentence ratings on the control values. It is used to quantify the rating difference of the generated sentences to an optimally controlled hypothetical. Through our human inspection of different control values in the library, we have a range of ratings that we wish to be controlled. This range is predefined for each attribute prior to our evaluation. The expected rating for each control value is therefore characterized by a linear interpolation of the minimum rating and maximum rating of the value. The error is defined by the absolute difference between the average rating of the sentences and the expected rating, then averaged over all the control values. For a given list of n average ratings $r_0, \ldots, r_{n-1}$ of sentences for each control value c, and the maximum and minimum range $R_{\text{max}}, R_{\text{min}}$, the Mean-MAE metric can be written as

$$\text{Mean-MAE} = \sum_{c=0}^{n-1} |r_c - r_c^*|,$$

$$\text{where, } r_c^* = R_{\text{min}} + \frac{c}{n-1} \times (R_{\text{max}} - R_{\text{min}}).$$

Mean-STD measures the variation of the sentence ratings on the control values. A good smooth

control method should be able to generate sentences of similar ratings. As the name suggests, this metric is calculated by averaging the standard deviations of ratings across different control values.

Relevance quantifies the utility of the responses in answering the questions. A perfect smooth control method should not sacrifice the utility for a smaller error or variation. Here, we employ `GPT-4-turbo` to judge the relevance between a question and a response. The specific prompt we use is provided in Appendix A.3.

5 Experiments Setup

In this section, we apply our proposed evaluation framework, along with the constructed benchmark dataset, to assess the smooth control capability of various modern LLMs through two viable training-free methods: **(1)** Prompting with semantic shifters, and **(2)** Editing the internal model representations. We first introduce the experiment settings and then present the results and analyses.

5.1 Baseline Methods For Smooth Control

Prompting LLMs. The most straightforward method to smoothly control the LLM to generate according to an attribute is to provide it with instruction on the degree level required. To achieve this, we need one description $\mathcal{D}_{\mathcal{A},cv}$ for each degree cv of the attribute $\mathcal{A}$:

$$\mathcal{R} = M_{\text{prompt}}(\mathcal{Q}, \mathcal{D}_{\mathcal{A},cv}),$$

We call this prompting method parameterized by the descriptions we choose. We consider two types of degree descriptions, first a list of semantic shifters that can describe the intensity paired with the adjective of the attribute (e.g., "a little bit angry" or "very angry"), and the second, a crafted list of phrases that not necessary sticks with a format (e.g., "slightly relaxed" or "extremely enraged"). The advantage of the first type is that they are seemingly easy to apply directly to different attributes, while for the second type, there is more flexibility in the descriptions. The exact descriptions we use for each attribute and the prompt templates to use these descriptions are in Appendix A.4 and A.7.

Representation Engineering (RepE). Different from prompting, RepE (Zou et al., 2023) is a top-down approach to post-processing pretrained models via manipulating their internal representations for understanding and controlling neural networks. Specifically, it involves two distinct steps in particular. (1) **Reading:** localizing the functional representations for a specific concept, which is generally achieved by analyzing the neural activities after stimulating the model with certain input prompts. The original stimulus prompts are manually written by humans, which have limited scope and lack generalizability to unseen concepts. In our experiments, we employ GPT-4 (OpenAI, 2023) to generate those stimuli automatically and the prompt template is in Appendix A.6. (2) **Controlling.** The extracted representations from the reading step are then utilized as high-dimensional vectors to perturb the original model representations to different extents indicated by a control strength, which perfectly aligns with the concept of control value in our task. Therefore, we specify the control strength for each control value of the attribute A. Such manipulation of the internal representations is also parameterized by the strength we indicate.

$$\mathcal{R} = M_{\text{RepE}}(\mathcal{Q}, \text{Strength}_{\mathcal{A},cv}),$$

5.2 Parameter Selection

It is not immediately clear whether the human-interpreted degree descriptions for prompting or the human-selected degree intensities in RepE transfer to a smooth degree control for the LLM. Therefore, we consider a "parameter selection" process for these two methods for calibration of the degrees. Specifically, we proactively consider a larger number of degree parameters (descriptions for prompting or strength for RepE), and obtain generations of the LLM based on the parameter through a held-out set of questions. Then, we select the sequence of parameters that leads to the best overall metric, which is defined and calculated as:

$$\text{Metric} = \frac{\text{Mean-MAE} + \text{Mean-STD}}{(R_{\text{max}} - R_{\text{min}}) * \text{Relevance}}$$

This metric is designed to determine a better set of generations from a specific smooth control method. The breakdown and intuitive explanations of this formula are as follows. **(1)** The nominator is the sum of the two aforementioned rating errors. A high Mean-MAE indicates misalignment with rating scales, while a high Mean-STD indicates unstable, varied ratings. To keep both values reasonable, we add rather than multiply them, as they share the same scale. Empirical evaluation shows that an unweighted average performs nearly best based on human inspection. The corresponding statistics

Method	Attr.	Mean-MAE (↓)						Mean-STD (↓)						Relevance (↑)					
		mistral 7b	llama 7b	llama 13b	llama 70b	gpt 3.5	gpt 4	mistral 7b	llama 7b	llama 13b	llama 70b	gpt 3.5	gpt 4	mistral 7b	llama-7b 7b	llama 13b	llama 70b	gpt 3.5	gpt 4
Prompt	A	0.87	0.63	0.72	1.50	0.63	0.51	0.83	0.97	0.68	0.87	0.93	1.03	0.79	0.79	0.91	0.83	1.00	0.99
	H	1.07	0.42	0.32	0.66	0.58	0.45	1.34	1.03	0.82	0.92	1.02	1.33	0.83	0.90	0.92	0.90	1.00	0.99
	F	1.33	1.08	1.24	1.21	1.14	0.73	1.47	1.19	0.90	1.10	1.06	1.07	0.75	0.97	1.00	0.93	0.96	0.99
	U	3.70	1.55	1.63	0.70	1.89	0.69	1.73	1.52	1.47	2.34	2.77	1.06	0.72	0.84	0.91	0.88	0.95	0.99
	C	1.30	1.11	1.30	0.73	2.36	0.79	3.59	3.62	2.68	4.94	1.62	3.41	0.76	0.95	0.93	0.90	0.98	1.00
RepE	A	1.94	1.33	1.32	-	-	-	0.87	1.22	1.13	-	-	-	0.81	0.93	0.95	-	-	-
	H	1.18	1.52	1.68	-	-	-	2.14	2.10	1.92	-	-	-	0.80	0.87	0.93	-	-	-
	F	1.86	2.06	1.90	-	-	-	1.44	1.50	1.59	-	-	-	0.93	0.82	0.94	-	-	-
	U	2.08	2.29	0.58	-	-	-	1.75	2.02	2.07	-	-	-	0.97	0.88	0.72	-	-	-
	C	1.26	1.40	1.39	-	-	-	0.92	0.76	0.90	-	-	-	0.95	0.97	0.67	-	-	-

Table 2: Evaluation results after parameter selection for each model and attribute. Here, 'Attr.' is short for 'Attribute', Mean-MAE denotes the calibration error, standard deviation indicates the robustness of smooth control, and relevance suggests if the generated response aligns with the topic. Some values are marked as '-' due to the constraints of accessing the model parameters and the coefficient range.

are exhibited in Appendix B. **(2)** The denominator is the multiplication of the normalization term and the relevance penalty factor. A low relevance score is undesirable, so we use its reciprocal to heavily penalize low-relevance generations.

The selection can be done efficiently by brute-force enumeration when the number of the total considered parameters is not too large and the specific number in our case is 20.

5.3 Experiment Settings

The evaluations are conducted on diverse LLMs for the smooth control of specific attributes. As such, we present the models, attributes, and datasets that are utilized in the experiments here.

Models. We employ both open-source and closed-source LLMs for our experiments. Specifically, we adopt Mistral (Jiang et al., 2023) and LLaMA2 (Touvron et al., 2023) at different scales for the experiments of editing the internal model representations, as it requires access to the model parameters. For prompting with semantic shifters, we further utilize GPT-3.5 (OpenAI, 2022) and GPT-4 (OpenAI, 2023) models.

Attribute. As explained in Section 3.2, we select Anger, Happiness, Formality, Understandability, and Conciseness as the attributes to evaluate. In particular, the intensity of Conciseness is measured differently than other attributes by directly counting the number of words in the responses.

Dataset. We adopt the constructed benchmark dataset introduced in Section 3.2, which consists of $1,500$ query sentences in total, with 500 for each of the aforementioned 5 attributes.

Metric. According to our evaluation framework introduced in Section 4, we adopt mean-MAE, standard deviation, and relevance as the main metrics.

6 Experiment results

6.1 Main Results

Table 2 shows the smooth control performance achieved by different models with different methods, on several attributes. One can observe that GPT-4 is significantly better than other models for all attributes, especially in terms of Mean-MAE, namely the consistency between the control values and the obtained attribute intensities. GPT-4 is also significantly better than other models in terms of the relevance between the model's response and the query, despite the potential cause being that GPT-4 is also used to evaluate such relevance.

Interestingly, we observe that model sizes may negatively affect the smooth performance. A relatively fair test bed for this is the Llama family, where one can observe that for most attributes, Mean-MAE decreases constantly as the model size increases from 7B, 13B, to 70B.

Finally, we also observe that prompting is almost as good as, if relatively better than repE. This implies that prompting is preferred in realistic applications of smooth control since it requires no access to the internal model representations and thus can be potentially applied to more LLMs.

6.2 Specificity of Parameter Selection in Intensity Calibration

In this section, we wish to explore whether the intensity descriptors we selected for each attribute and each model are specific to the model or the attribute. Here we mainly conduct the investigation based on prompting since prompting is more preferred than RepE based on the above results.

Should intensity descriptors be specific to attribute? We validate whether it is possible to use a universal set of descriptors to control the inten-

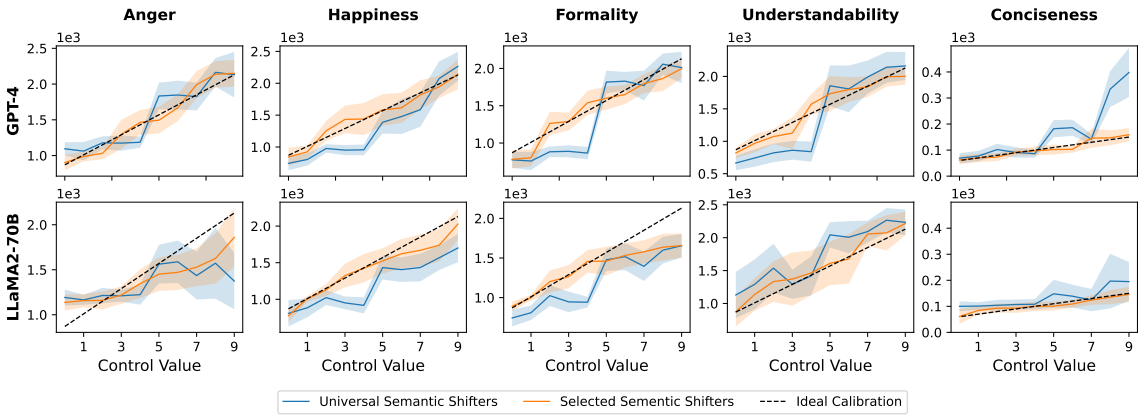

Figure 4: Comparisons between prompting with universal and selected semantic shifters. The Y axis is the attribute intensity. The black dashed lines are the ideal correlation between the control value and the attribute intensity.

	Mean-MAE (↓)						Mean-STD (↓)						Relevance (↑)					
	mistral 7b	llama 7b	llama 13b	llama 70b	gpt 3.5	gpt 4	mistral 7b	llama 7b	llama 13b	llama 70b	gpt 3.5	gpt 4	mistral 7b	llama-7b 7b	llama 13b	llama 70b	gpt 3.5	gpt 4
A	1.94	1.08	0.84	1.49	0.86	0.85	1.17	1.10	1.20	1.67	1.23	1.47	0.78	0.75	0.86	0.72	0.92	0.99
H	1.26	0.51	0.50	0.96	0.77	1.25	1.48	1.25	1.23	1.23	1.42	1.60	0.81	0.87	0.94	0.87	1.00	0.99
F	1.46	1.48	1.54	1.50	1.36	1.29	1.19	1.23	1.03	1.23	1.16	1.12	0.75	0.95	0.99	0.91	0.96	0.99
U	3.35	1.65	1.70	2.68	4.03	2.30	1.67	1.56	1.63	2.08	2.09	1.12	0.45	0.81	0.83	0.91	0.86	0.93
C	1.60	1.08	1.30	0.90	3.19	1.63	2.66	2.71	1.98	2.43	1.36	2.13	0.77	0.92	0.93	0.91	0.89	1.00

Table 3: Prompting with intensity descriptors that are not specific to models.

extremely not	a little bit
very not	somewhat
moderately not	moderately
somewhat not	very
a little bit not	extremely

Table 4: Universal Semantic Shifters

Attri	Mean-MAE		Mean-STD		Relevance	
	gpt-4	llama-70b	gpt-4	llama-70b	gpt-4	llama-70B
A	1.00	2.02	1.27	1.45	0.98	0.77
H	1.76	2.29	1.31	1.18	0.97	0.83
F	1.75	2.3	1.03	1.06	0.99	0.95
U	2.01	1.84	1.64	2.07	1.00	0.83
C	7.17	3.07	3.81	4.61	1.00	0.86

Table 5: Baseline with fixed semantic shifters.

sities of all attributes listed. If possible, such a set can greatly ease the implementation of smooth control of LLMs and reduce the inference cost for selecting specific descriptors of each attribute.

To this end, we experiment with using a set of fixed semantic shifters to modulate the intensity of an attribute in prompting. In specific, we prompt GPT-4 multiple times to generate 30 different adverbs of degrees that are commonly used. We then select 10 that appear the most frequently in responses, which are shown in Table 4.

Figure 4 and Table 5 show the results of using fixed semantic shifters for prompting LLMs. We observe that such fixed semantic shifters achieve significantly worse performance in smooth control, especially in terms of Mean-MAE. This means fixed semantic shifters cannot properly control the attribute to match the desired intensities.

Are intensity descriptors specific to model? Among our experiment results, we found that the intensity descriptors selected to achieve the best smooth control performance vary significantly across models. According to our observation of the results for attribute "Formality", to achieve an intensity level of 3, GPT-4 would prefer "Highly Inappropriate" in its prompt. In contrast, Llama2-70b would prefer "Neural" to achieve the same intensity level. Further, we found that the intensity descriptors preferred by different models may not even be consistent in terms of order.

Since the best intensity descriptors are not specific to the model, one cannot simply transfer the intensity descriptors selected for one model to another model. We conduct an additional experiment to demonstrate this. In Table 3, for each model, we use all models except this model to select the intensity descriptors. One may observe that these intensity descriptors transferred from other models cause significantly worse smooth control performance, especially in terms of Mean-MAE. This shows that it is necessary to select intensity descriptors specific to each model to properly control the attribute to match the desired intensity.

7 Conclusions and Future Work

This work studies smoothly controllable text generation for large language models. We created an evaluation system on five different attributes to evaluate smooth control methods on different intensity levels for three metrics: error, variation of the generated sentence's intensities and the relevance to the generation questions. The system is implemented based on Elo ratings, automatically evaluating using LLMs, and designed to be efficient in evaluation. We evaluate two representative methods, prompting and representation engineering. We find that (1) Model sizes may negatively affect the smooth performance. (2) Prompting is almost as good as, if relatively better than repE.

Limitations

Our work presents an evaluation of smooth control methods for LLM generations. There are several limitations that we have considered:

- We used GPT-4 as an automatic evaluator in building our evaluation system, mainly for reducing human effort. While we have verified its closeness with human preference on all the 5 attributes we considered, we admit that our system will suffer from the same limitations of using LLM as annotators, such as not being robust to certain (manually crafted) sentences, and not being a free service to use, especially that we find not as competent LLMs (e.g., GPT-3.5) do not have a similar strong annotation power.
- We mainly evaluated two training-free methods, Prompting and Representation Engineering for their soft control ability, due to their simplicity and representativeness. Other soft control methods, including some that require model fine-tuning, could be evaluated in future work.

Acknowledgement

We thank the anonymous reviewers for their insightful comments. Our work is sponsored in part by NSF CAREER Award 2239440, NSF Proto-OKN Award 2333790, as well as generous gifts from Google, Adobe, and Teradata. Any opinions, findings, conclusions, or recommendations expressed herein are those of the authors and should not be interpreted as necessarily representing the views, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright annotation hereon.

References

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *ArXiv*, abs/1912.02164.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *International Conference on Machine Learning*.

Chenyang Huang, Osmar R Zaiane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic dialogue generation with expressed emotions. In *North American Chapter of the Association for Computational Linguistics*.

Qiushi Huang, Yu Zhang, Tom Ko, Xubo Liu, Boyong Wu, Wenwu Wang, and Lilian Tang. 2022. Personalized dialogue generation with persona-adaptive attention. *ArXiv*, abs/2210.15088.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *ArXiv preprint*, abs/2310.06825.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020. Hooks in the headline: Learning to generate headlines with controlled styles. In *Annual Meeting of the Association for Computational Linguistics*.

Zhijing Jin, Di Jin, Jonas W. Mueller, Nicholas Matthews, and Enrico Santus. 2019. Imat: Unsupervised text attribute transfer via iterative matching and translation. In *Conference on Empirical Methods in Natural Language Processing*.

Muhammad Khalifa, Hady ElSahar, and Marc Dymetman. 2020. A distributional approach to controlled text generation. *ArXiv*, abs/2012.11635.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq R. Joty, Richard Socher, and Nazneen Rajani. 2020. Gedi: Generative discriminator guided sequence generation. In *Conference on Empirical Methods in Natural Language Processing*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *North American Chapter of the Association for Computational Linguistics*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *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)*, abs/2101.00190.

Qian Liu, Yihong Chen, B. Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020a. You impress me: Dialogue generation via mutual persona perception. In *Annual Meeting of the Association for Computational Linguistics*.

Ruibo Liu, Guangxuan Xu, Chenyan Jia, Weicheng Ma, Lili Wang, and Soroush Vosoughi. 2020b. Data boost: Text data augmentation through reinforcement learning guided conditional generation. In *Conference on Empirical Methods in Natural Language Processing*.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. Powertransformer: Unsupervised controllable revision for biased language correction. In *Conference on Empirical Methods in Natural Language Processing*.

Andrea Madotto, Zhaojiang Lin, Yejin Bang, and Pascale Fung. 2020. The adapter-bot: All-in-one controllable conversational model. In *AAAI Conference on Artificial Intelligence*.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`.

OpenAI. 2023. Gpt-4 technical report.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2019. Automatically neutralizing subjective bias in text. *ArXiv*, abs/1911.09709.

Sudha Rao and Joel R. Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *North American Chapter of the Association for Computational Linguistics*.

Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating summaries with controllable readability levels. In *Conference on Empirical Methods in Natural Language Processing*.

Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych. 2021. Structural adapters in pretrained language models for amr-to-text generation. *ArXiv*, abs/2103.09120.

Tianxiao Shen, Tao Lei, Regina Barzilay, and T. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *ArXiv*, abs/1705.09655.

Haoyu Song, Yan Wang, Kaiyan Zhang, Weinan Zhang, and Ting Liu. 2021. Bob: Bert over bert for training persona-based dialogue models from limited personalized data. *ArXiv*, abs/2106.06169.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating responses with a specific emotion in dialog. In *Annual Meeting of the Association for Computational Linguistics*.

Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J. Martin, Animesh Mehta, Brent Harrison, and Mark O. Riedl. 2018. Controllable neural story plot generation via reward shaping. In *International Joint Conference on Artificial Intelligence*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *ArXiv*, abs/1901.08149.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. Controllable story generation with external knowledge using large-scale language models. In *Conference on Empirical Methods in Natural Language Processing*.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2022. Tailor: A prompt-based approach to attribute-based controlled text generation. *ArXiv*, abs/2204.13362.

Yoel Zeldes, Dan Padnos, Or Sharir, and Barak Peleg. 2020. Technical report: Auxiliary tuning and its application to conditional text generation. *ArXiv*, abs/2006.16823.

Yizhe Zhang, Guoyin Wang, Chunyuan Li, Zhe Gan, Chris Brockett, and Bill Dolan. 2020. Pointer: Constrained progressive text generation via insertion-based generative pre-training. In *Conference on Empirical Methods in Natural Language Processing*.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. Style transfer as unsupervised machine translation. *ArXiv*, abs/1808.07894.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Yinhe Zheng, Rongsheng Zhang, Xiaoxi Mao, and Minlie Huang. 2019. A pre-training based personalized dialogue generation model with persona-sparse data. In *AAAI Conference on Artificial Intelligence*.

Daniel M. Ziegler, Nisan Stiennon, Jeff Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *ArXiv*, abs/1909.08593.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

Appendices

A Prompt Templates

We list all prompt templates we used in this paper.

A.1 Question Generation

Our dataset consists of questions that potentially can be answered with different degrees of different attributes. The template to generate the questions is

```
Generate 10 prompts that can be answered with
varying degrees of <concept>.
```

A.2 Pairwise Annotation

This template is used to compare two responses to decide which shows a greater degree of the concept.

```
For each pair of responses, identify which
response expresses more <concept>.
Write the pair number followed by '1' if the
first response is more <concept>, or '2' if the
second response is more <concept>.
Format your response like this: '1. 1', '2. 2',
etc.
```

A.3 Relevance Annotation

This template is used to Judge if a response is relevant (1) or not (0) to the query.

```
Given the following query and response, please
assess whether the response is relevant to the
query. Answer with '1' if the response is
relevant, and '0' if it is not relevant.
```

A.4 Prompting with Degree Descriptions

This template is used to respond to queries with a specified emotional tone or style.

```
Please respond to {{queries[i]}} with a
paragraph in a [tone | style] that is
{{semantic shifter}}. The response should be
three sentences long.
```

A.5 Generating Degree Descriptions

This template is used to identify words or phrases that can shift the meaning of a concept, either intensifying or diminishing its strength.

```
Describing <concept> levels on a scale from -9
to 10 using phrases.
```

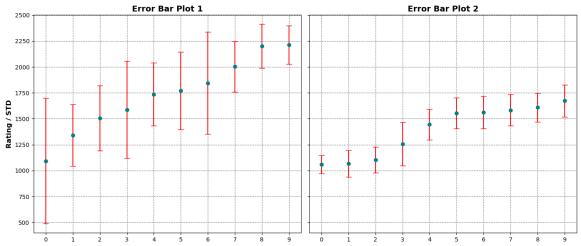

Figure 5: Examples for human evaluation.

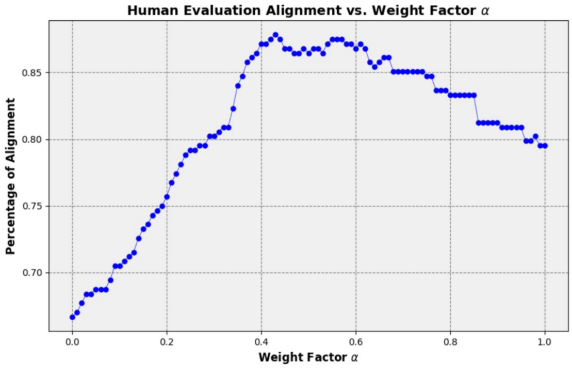

Figure 6: Alignment with humans for different weight factors.

A.6 Stimulus Prompts Generation.

```
Generate 10 prompts that can stimulate
<concept>.
```

A.7 Candidates for Semantic Shifters

B Parameter Selection Analysis

We considered different sets of α (from 0 to 1) for the weighted average of Mean-MAE and Mean-STD to calculate the overall metric.

$$\alpha \times \text{Mean-MAE} + (1 - \alpha) \times \text{Mean-STD} \quad (1)$$

For each weight factor α, we considered pairs of error bar plots of the average and standard deviation values, and asked humans to judge which plot is better as shown in Figure 5. We compare the human evaluation result with the result that our metric provides, and record the percentage of alignment. As shown in Figure 6, alignment follows a bell curve, peaking at 0.87 when α is 0.5-ish. Therefore, we directly adopt the vanilla average of these two rating errors rather than the weighted ones.

C Generated Data Examples

Attribute	Query
Anger	Why did you use my personal items without asking for permission? What exactly is causing the continuous delays on the construction project you assured would be finished by now? What possessed you to share our private conversations with others? What makes you think it's okay to leave such a mess in the kitchen every time you cook? Why do you consistently interrupt me when I'm trying to have a serious conversation?
Happiness	Can you share a story of a time when you overcame a challenge and felt great contentment? What activity do you enjoy doing that always lifts your spirits? What is your favorite memory from childhood that still makes you smile? What was your reaction when you received the gift you've always wanted? Imagine your perfect day; what does it look like and how does it make you feel?
Formality	Could you please tell me about the latest developments in artificial intelligence? Describe the process of photosynthesis in plants. What steps should I take to prepare for a job interview? How does one go about making a traditional Italian pizza from scratch? Describe the impact of social media on interpersonal communication in today's society.
Understandability	Illustrate the impact of climate change on polar bear habitats. How does a computer process information? Assess the philosophical implications of artificial intelligence in society. Outline the basic tenets of existentialist philosophy. What are black holes, and can they affect our planet?
Conciseness	What steps would one take to secure their home Wi-Fi network? Elucidate the reasons behind the diversity of languages spoken around the world. Can you provide the steps involved in solving a Rubik's Cube? Describe the roles enzymes play in the human digestive system. Explain the theory of relativity and its implications for physics.

Table 6: Benchmark data examples generated by GPT-4

Bin ID	Rating	Example Sentence
0	871	I hold a deep belief in the natural flow of truth, as it always finds a way into the open, much like a gentle stream that carves its path through the landscape. In taking action, my intention was to nurture understanding without disruption, trusting that, with time, our paths would align in clarity and harmony. I welcome this moment for reflection and connection, serene in the knowledge that openness will guide us toward mutual tranquility.
1	1027	I appreciate your patience and understanding, as time often seems to flow like a gentle stream, unfettered by our human constructs, leading me to drift subtly off schedule. Each moment unfolds with its own unique rhythm and sometimes that melody harmonizes with the clock differently than intended. Rest assured, my intentions are to honor our appointments just as I honor each serene pulse of the present, and I am taking measures to align my peace with punctuality.
2	1147	I deeply appreciate your patience and understanding as we address the timing concerns. My tardiness has been primarily a result of unexpected complications that arise despite my initial planning and efforts to arrive on time, which I'm earnestly working to overcome. Rest assured, I value our meetings and am implementing new strategies to ensure that I honor our appointments punctually moving forward.
3	1302	Honestly, I didn't realize I had left the kitchen as such a disaster; clearly, cooking got ahead of me this time. Nevertheless, it's frustrating to hear that my oversight has caused inconvenience since I usually tidy up after myself. I'll address the mess immediately, as I certainly didn't mean to add any stress to our day.
4	1441	I assure you that I am fully aware of the importance of taking responsibility for my own actions. Your implication that I habitually push blame onto others is neither fair nor accurate. However, I'll reflect on this feedback and commit to being more mindful of how I address issues in the future.
5	1567	Frankly, I can't fathom what was going through the mind of the person who brazenly spoiled the movie's ending for everyone. It's basic movie-watching etiquette to keep plot twists to oneself, especially in a communal space where the anticipation is part of the communal experience. Some courtesy would be appreciated, to not ruin the suspense we've all been patiently waiting to enjoy together.
6	1694	I am growing increasingly frustrated by the lack of updates regarding the refund that was due two weeks ago. Your inability to process it in the promised timeframe is both inconvenient and unacceptable. I need a clear explanation for this delay and an immediate resolution to ensure I receive my refund forthwith.
7	1873	Seriously, the nerve of some people cutting in line as if the concept of waiting their turn simply evaporates when it comes to them! It's a blatant disregard for common courtesy and the unspoken social contract we all agree to when joining a queue. Their sense of entitlement is astounding and a slap in the face to everyone who respects the order of things.
8	1995	I've had enough of constantly being painted as the one who avoids accountability! Frankly, it's exhausting and hypocritical for you to suggest I haven't faced my own faults when you've hardly glanced at your own missteps. It's high time for a reality check on both ends because I refuse to be the scapegoat for problems I haven't caused alone!
9	2168	Absolutely unbelievable, isn't it? I'm constantly stuck picking up the pieces after your careless blunders, pouring my energy into fixing what should never have been an issue in the first place! I won't stand for this any longer; it's about time you step up and take responsibility for your own actions rather than expecting me to clean up your incessant disasters!

Table 7: Responses with different intensities in the attribute of anger.

Anger	Happiness	Formality	Understandability	Conciseness
serenely peaceful	despair	extremely disrespectful	completely unintelligible	extremely redundant
deeply relaxed	miserable	highly inappropriate	extremely confusing	highly redundant
very calm	very unhappy	very casual	very hard to understand	very redundant
quite tranquil	unhappy	informal	quite difficult to understand	quite redundant
mildly peaceful	slightly unhappy	casual	challenging to understand	moderately redundant
slightly relaxed	neutral/negative	slightly casual	hard to follow	slightly redundant
neutral, neither calm nor angry	neutral	neutral	somewhat unclear	marginally wordy
slightly irritated	neutral/positive	slightly formal	slightly unclear	somewhat wordy
mildly annoyed	slightly happy	moderately formal	almost clear	mildly wordy
neutral, balanced emotion	content	formal	neutral	neutral
slightly upset	satisfied	very formal	fairly easy to understand	mildly concise
moderately annoyed	cheerful	highly formal	clear	somewhat concise
fairly irritated	happy	ceremonial	very clear	moderately concise
quite angry	very happy	old-fashioned formal	extremely clear	quite concise
very angry	joyful	courtly	crystal clear	very concise
intensely furious	elated	aristocratic	intuitively understandable	highly concise
extremely enraged	overjoyed	regal	effortlessly understandable	extremely concise
seething with rage	ecstatic	imperial	instantly understandable	terse
nearly uncontrollable anger	blissful	divine	universally understandable	overly terse
utterly livid, maximum anger	nirvana	transcendent	absolute clarity	cryptic

Table 8: Candidates for Semantic Shifters