From Black Box to Controller: Steering LM Behavior via Sparse Autoencoder Activations

Anonymous ACL submission

Abstract

Controlling the behavior of the language model (LM) during inference—such as adjusting toxicity, sentiment tendency, and degree of politeness—is crucial for natural language processing. In this work, we introduce **NeuroSteer**, a plug-and-play framework that facilitates the adjustment of the LM behavior without domain-specific training. NeuroSteer leverages a Sparse AutoEncoder (SAE) as an output controller, activating SAE neurons linked to target behaviors, extracting the corresponding feature residuals, and adding them to the model's hidden states to directly influence the generation process. This feature-space intervention amplifies the weight of target features in the latent representations, enabling precise control over the model's output distribution. NeuroSteer effectively alters the LM's stance, sentiment, toxicity, and politeness during inference, achieving SOTA performance across four datasets while maintaining a balance between generation quality and behavioral adjustments. Unlike fine-tuning, NeuroSteer enables fast domain adaptation by calculating activations on hundreds of examples in seconds, without the need for retraining. Furthermore, our work not only provides a possible task adaptation solution, but layer-wise interventions also provide deeper insights into the model's mechanisms, shedding light on how concepts are represented in the LM and how combining feature vectors influences behavior. Code: `http://acl.anonymous-demo.fun`

1 Introduction

Task alignment and behavior adaptation remain challenges for the language model (LM). For instance, in medical scenarios, models must avoid toxic or misaligned outputs, whereas in customer service, adjusting to suppress negative emotions and enhance politeness is critical. Current alignment techniques like full fine-tuning (Devlin, 2018), LoRA (Hu et al., 2021), and RLHF (Ouyang et al., 2022) require substantial resources, including parameter adjustments, large datasets, computing, etc. Moreover, they are not flexible, as each scenario's alignment requires retraining for specific tasks.

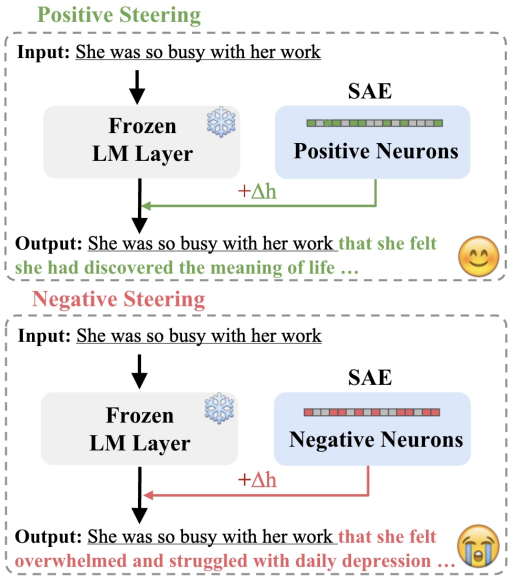

Figure 1: Overview of sentiment steering. The LM generates positive sentiment text steered by positive SAE activations Δh. Vice versa.

Therefore, more flexible intervention methods are needed, and activation engineering provides new solutions. Studies (Jiang et al., 2024; Hernandez et al., 2024; Park et al., 2024) show that LM encodes knowledge along linear directions in hidden activations, enabling latent edits without parameter changes. To change model output, activation engineering methods (Elhage et al., 2022; Li et al., 2024; Panickssery et al., 2024) use the steering vector to shift LM activation to a linear direction. Although these methods can change LM output with some promise, they encounter two critical limitations: (1) a single vector restricts the scope and capability of controlling complex LM behavior. (2) calculating the mean difference vector

Methods	No Train	Quantifiable	Plug-and-Play	No Inference Delay	Combined Steering (§D)	Interpretability
Fine-Tuning(2018)	✗ (Per Task)	✗	✗	✓	✗	✗
LoRA(2021)	✗ (Per Task)	✗	✗	✓	✗	✗
ITI(2024)	✓	✓	✓	✓	✗	✗
CAA(2024)	✓	✓	✓	✓	✗	✗
LM-Steer(2024)	✗ (Per Task)	✓	✗	✗	✗	✗
NeuroSteer	○ Only Once	✓	✓	✓	✓	✓

Table 1: NeuroSteer demonstrates advantages in aspects such as computational efficiency and flexibility.

introduces noise, leading to undesired outputs.

To address these limitations, we propose NeuroSteer, which controls LM behaviors by combining a set of basic feature vectors in the hidden state. We introduce Sparse Autoencoder (SAE) as LM's controller to collect and combine these vectors.

To steer the LM toward generating positive output, we activate and combine SAE neurons corresponding to basic positive features, such as "!", "amazing" among many others. As shown in Figure 1, the SAE decoder reconstructs these neuron activations into a set of feature vectors Δh, which are then added as residuals to the LM activations. Residuals are plug-and-play, as it requires no weight editing of the model. During inference, the residual addition influences the model probability distribution on all next-token predictions, steering the LM output toward the desired behavior.

To select neurons representing positive features, we use a frequency filter on comparative data. Neurons that are frequently activated in positive samples but not in negative ones are key to positive behaviors. Neurons activated in both samples represent noise and are not activated for steering.

NeuroSteer advantages shown in Table 1:

- **Train Once, Cross Tasks:** Pretrained SAE can handle cross-task interventions without the need for retraining.
- **Compute Effectively:** Achieves better control than LoRA with 14% training time and 10% domain data (§4.5).
- **Quantifiable:** NeuroSteer can adjust intervention intensity and concept hierarchy, enabling precise control from lexical (e.g., "against") to behavioral (e.g., stance) levels (§8).
- **Plug-and-Play:** Our intervention can be dynamically plug-in, modified, or combined. It works without inference delay.

Our contributions can be summarized as follows:

- We develop NeuroSteer, an LM control framework introducing SAE feature vector control to LM behavior precisely.
- NeuroSteer achieves SOTA results on sentiment, toxicity, stance, and politeness steering tasks while excelling in computing efficiency, quantifiability, and flexibility.
- Deeper insights are gained into how concepts are represented in LM and how combining basic feature vectors affects the LM's behavior.

2 Related Work

The work of NeuroSteer is inspired by the concept of linear representations and aims to address issues such as noise in current intervention methods.

2.1 Linear Interventions

Although LM's hidden states are often seen as opaque and chaotic, studies show they develop semantically meaningful internal structures. Jiang et al. (2024) and Li et al. (2024) suggest that semantic concepts are generally encoded linearly in the hidden representation space of LMs. For instance, concepts such as sentiment polarity (Tigges et al., 2023), time, and space (Gurnee and Tegmark, 2024) can be represented through linear directions in hidden states. This makes it possible to change LM output by editing hidden states.

Following the idea of linear representations, approaches like Panickssery et al. (2024) compute mean hidden state difference activations along the vector connecting the means of the true and false distributions, employing the mean difference vector to improve the truthfulness of LM. Similarly, Dong et al. (2023) and Tigges et al. (2023) steer the LM latent state with steering vectors using SVD decomposition. Most activation engineering methods (Yunfan et al., 2025; Yang et al., 2025; Stickland et al., 2024) calculate mean differences between positive and negative samples, applying a single steering vector directly on LM hidden states.

2.2 Limitations of Linear Interventions

Despite their potential, linear interventions face two critical challenges rooted in the fundamental properties of LM representations.

Feature Superposition as Noise Amplifier: The dense, highly superimposed nature of LM hidden layers causes unintended interference during interventions. As shown by Elhage et al. (2022), individual neurons rarely encode single features – GPT-2's 768-dimensional latent space must encode over 10^4 distinct features (Gurnee et al., 2024). This superposition manifests practically: Amplifying the steering vector results in unintended token repetitions (e.g., "great great great...") in most cases, indicating that noise features are being activated along the steering direction.

Single-Vector Control Bottleneck: The mean single difference vector approach (Stickland et al., 2024) limits intervention flexibility. Attempts to scale intervention strength by multiplying the vector often lead to nonlinear effects. Furthermore, composite behaviors requiring multiple feature combinations (e.g., "truthful + formal style") cannot be systematically constructed.

2.3 SAE Acts as Superposition Resolvers

Recent advances in interpretability provide a path forward through SAE. Originally developed for feature disentanglement (Templeton et al., 2024), SAE impose sparsity constraints on high-dimensional latent spaces (typically 10-100 times larger than original embeddings). This forces neurons to activate independently for distinct semantic features. For example, in GPT-2 SAEs, specific neurons fire exclusively for concepts like "email addresses" or "chemical formulas" (Kissane et al., 2024).

The SAE architecture naturally separates superimposed features: Given an encoded representation on LM, the encoder produces sparse features, where only a small subset of neurons are activated, while the decoder reconstructs the original encoded representation. Crucially, the overcomplete basis enables SAE neurons to remain relatively independent and non-interfering, suppressing off-target noise. Independent Neurons make SAE promising mediators for interventions, activating the relatively independent neurons in the SAE enables the targeting of specific features while avoiding unintended activations.

3 Method

To address the noise issues introduced by existing mean difference intervention methods in Section 2, we propose the two hypotheses to denoise.

3.1 Preliminary

Hypothesis 1 (Compositional Feature Control)**.** *LM behaviors can be controlled through linear combinations of basic semantic features.*

To resolve superimposed noise, we combine multiple independent feature vectors to intervene in the LM, operating within a relatively clean sparse space. The decoder of the SAE serves as the perfect combiner for these basic feature vectors. In sparse space, SAE neurons encode mutually independent features. By selectively activating neurons corresponding to specific features while masking out noise, SAE can achieve precise control over LM behaviors.

Furthermore, we introduce the Frequency-Based Denoising hypothesis to reduce the noise from undesired features in small-scale datasets.

Hypothesis 2 (Frequency-Based Denoising)**.** *Feature importance follows activation statistics*

High-level behaviors exhibit statistical patterns in large contrastive datasets which are also reflected in activation statistics. Neurons that are activated consistently in positive samples (e.g., happy words) but rarely in negative samples represent the desired behaviors, while neurons that are frequently co-activated in both positive and negative samples are indicative of noise (e.g., conjunctions).

3.2 Method Overview

As illustrated in Figure 2, NeuroSteer learns high-level behavioral transformations directly from large-scale positive and negative tokens. First, it collects GPT token activations for both sample types, encodes them through an SAE encoder, and calculates SAE neuron activation frequencies. Second, it selects positive SAE neurons based on those that are activated frequently in positive tokens and not in negative tokens. Finally, it activates these neuron combinations, decodes them into GPT's hidden space through residual connections Δh, and steers model behavior toward positive directions.

3.3 SAE Neuron Selection

Following hypothese 2, we filter positive neurons on large-scale contrastive data. To select the target neurons for steering, GPT activations on the contrastive texts are sampled to obtain frequency SAE statistics.

GPT Hidden State Collection: First, we input the positive and negative texts into the GPT model and collect the hidden states for each token. For

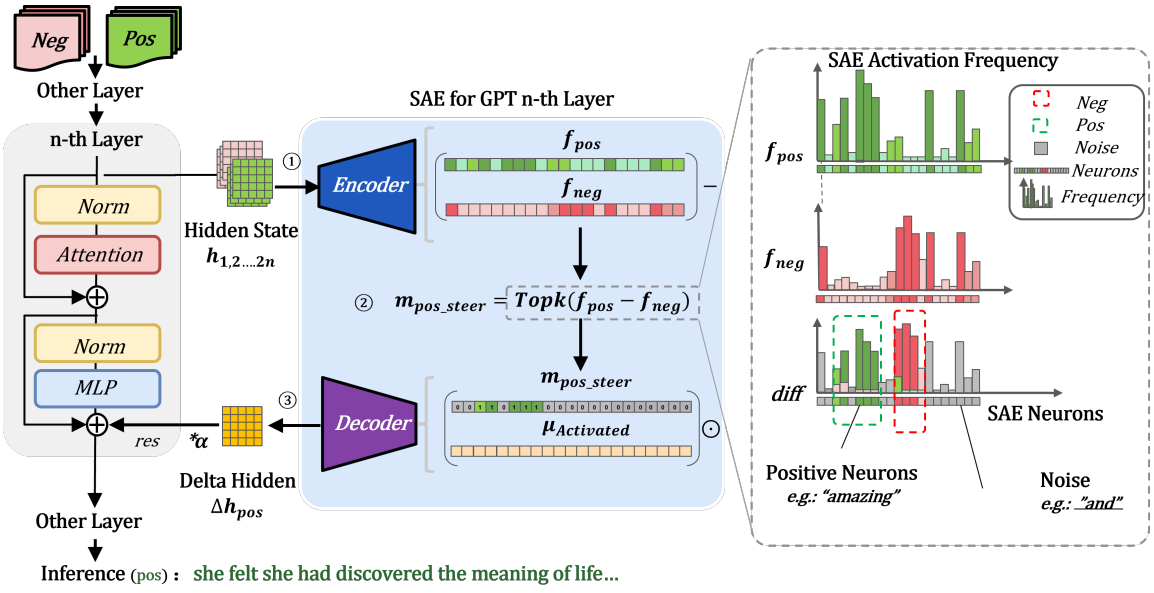

Figure 2: An overview of the NeuroSteer architecture. Data interaction between GPT and SAE occurs in three steps: ① **Neuron Selection**: The GPT collects neuron activation frequencies of positive and negative samples; ② **Denoising**: The SAE extracts target neurons and denoises based on frequency differences; ③ **Intervention Generation**: The SAE activates target neuron combinations and decodes their effects into GPT hidden space.

positive sentiment texts, n_{pos} tokens are used, and for negative counterparts, n_{neg} tokens are used.

SAE Encoding: For each token hidden state $\mathbf{h}^i_{\text{GPT}} \in \mathbb{R}^d$, the SAE encoder maps it to a sparse representation in higher-dimensional space :

$$\mathbf{h}^i_{\text{SAE}} = \text{Enc}\left(\mathbf{h}^i_{\text{GPT}}\right), \quad \mathbb{R}^d \to \mathbb{R}^k, \qquad (1)$$

where $k \gg d$ enables sparse feature separation through the SAE's dimensional expansion.

Frequency Statistics: For each SAE neuron $j \in \{1, ..., k\}$, we compute activation frequencies across positive $f_{\text{pos},j}$ and negative samples $f_{\text{neg},j}$:

$$f_{\text{pos},j} = \frac{1}{n_{\text{pos}}} \sum_{i=1}^{n_{\text{pos}}} \mathbb{I}(\mathbf{h}^i_{\text{SAE},j} > 0) \in \mathbb{R}^k, \qquad (2)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

3.4 Denoising

Figure 2(c) shows that green neurons activate for positive samples, while red neurons represent activations for negative samples, with color intensity indicating activation frequency. According to hypothesis 2, neurons encoding positive features exhibit higher activation frequency in positive samples. To filter out the noise and select the TopK neurons for target steering, the difference in activation frequencies between positive and negative samples is computed, as illustrated in Equation 3.

$$m_{\text{pos\_steer},j} = \text{TopK}(f_{\text{pos},j} - f_{\text{neg},j}) \in \mathbb{R}^k, \quad (3)$$

where $f_{\text{pos},j}$ and $f_{\text{neg},j}$ represent the activation frequencies of the j-th neuron for positive and negative samples, respectively. $m_{\text{pos\_steer},j}$ denotes the selected neurons after applying the Top-K ranking based on the activation frequency difference, and K is the number of neurons selected for target steering.

3.5 Intervention Generation

Following Hypothesis 1, we construct steering interventions through sparse neuron combinations:

Manual Activation: We will activate these target neurons based on the mean activation value $\boldsymbol{\mu}_j$ of each neuron j, then perform a dot product operation between this value and the target neuron 0-1 position mask $m_{\text{pos\_steer},j}$.

SAE Decoding: Sparse steering features are projected back to GPT latent space via SAE decoder:

$$\Delta h = Dec(\mathbf{m}_{\text{pos\_steer},j} \odot \boldsymbol{\mu}_j), \quad \mathbb{R}^k \to \mathbb{R}^d, \quad (4)$$

where SAE Decoder $\text{Dec}(\cdot)$ maintains reconstruction fidelity through pretrained decoding weights.

Inference Steering: During inference, the steering hidden state Δh is injected into GPT residual connection for each token, as shown in Equation 5:

$$h'_{\text{GPT}} = h_{\text{GPT}} + \alpha * \Delta h \in \mathbb{R}^d, \qquad (5)$$

The intensity parameter α controls intervention strength while maintaining linguistic coherence.

Transfer to Other Tasks: This section demonstrates NeuroSteer's output steering capability through emotional polarity reversal in GPT-2, transforming positive sentiment to negative. The steering mechanism extends to toxicity mitigation via adversarial data substitution between toxic and non-toxic samples. Further generalization enables stance modification and politeness adjustment, as empirically validated in our experiments.

4 Experiment

To verify the effectiveness of NeuroSteer, we conducted rigorous experiments on four open-ended tasks, that are **Sentiment Steering**, **Detoxify**, **Stance Steering** and **Politeness Steering**. All experiments followed a fundamentally similar approach: first, extracting SAE neurons from data with different polarities (e.g., texts with positive and negative sentiments), and then calculating the Δh for steering. After generating the target polarity steering texts, a classifier is used to classify the polarity of all generated texts. The proportion of texts with the target polarity steering is then calculated. This proportion serves as a measure of the steering capability of the method, with a higher change indicating stronger steering ability.

For sparsity, we use GPT-2-small (Radford et al., 2019) and Gemma-2-2B[1] as base models, with a latent dimension of 768, and intervene at the 6th layer. Additionally, we apply a pre-trained SAE with a larger latent dimension of k (Kissane et al., 2024). It provides the sparsity basis required by hypothesis1, with feature separation while maintaining reconstruction fidelity. Some NeuroSteer parameters ($k = 100$, $layer = 6$) are fixed.

4.1 Experimental Tasks

The four tasks can be categorized into two groups: **typical tasks** and **innovative tasks**. The sentiment steering and detoxify tasks are typical tasks, for which numerous previous works have provided baselines, making them common experimental tasks for evaluating steering methods. The stance and politeness steering tasks are innovative tasks, that have been scarcely explored in previous literature, and we provide baseline results for our method on these two tasks.

Sentiment Steering: In this typical task, we calculate and select the steering neurons on 3452 data samples from Stanford Sentiment Treebank (SST5, Socher et al., 2013) for sentiment steering. SST5 samples are classified for positive or negative sentiment and are used to compute positive and negative frequency differences for targeted neuron selection. For sentiment evaluation, we assess the model using the Open WebTextCorpus, an out-of-domain dataset with 2500 prompts from an OpenWeb sentiment classifier (Hartmann et al., 2023).

Detoxify: By controlling the toxic generation, Neurosteer can reduce harmful or offensive content. Jigsaw toxicity classification dataset (cjadams et al., 2017) with 155k data samples is a very suitable dataset for calculating Δh. For out-of-domain evaluation, we use a random sample of 10K non-toxic prompts from the RealToxicity Prompts dataset (Gehman et al., 2020) to assess toxicity reduction in GPT-2 generation.

Stance Steering: Stance detection determines whether the speaker is in favor of, against, or neutral to a target event. After steering, GPT can reverse the originally agreed-upon stance in the latter part of a sentence. We test stance reversal on the StanceSentences dataset (Reyes, 2024), which contains 972 samples, using the training and test sets of it to compute the feature vectors Δh and evaluate the results.

Politeness Steering: Controlling politeness in language models ensures that the generated outputs align with expected social norms and etiquette. The politeness control experiment follows the same approach as stance control. In the politeness steering experiment, we use huggingface politeness corpus (2024) with 10.9k samples for training and testing, with a 3:1 train-test split.

4.2 Steering Baselines

In two typical experiments, baselines can be divided into two categories depending on whether training is required.

Trained Methods include **DAPT** (Gururangan et al., 2020) and **PromptT5** (Raffel et al., 2020), which pretrain the LM on a non-toxic subset of OpenWebText (filtered via the Perspective API), and **PPLM** (Dathathri et al., 2019), which uses gradients from sentiment and toxicity classifiers to update the LM's hidden representations. Additionally, **LM-Steer** (Han et al., 2024) freezes the GPT layers and adds a fine-tuning layer.

Plug-and-Play Methods include **GeDi** (Krause et al., 2020), which uses Bayesian rules for class-conditioned generation; **MuCoLa** (Kumar et al., 2022), which formulates text generation as an op-

[1]See Appendix C for cross-model validation

Target	Method	Sentiment Positivity/%			Fluency	Diversity		
		Positive prompts	Neutral prompts	Negative prompts	Output ppl. ↓	Dist-1↑	Dist-2↑	Dist-3↑
Positive ↑	NeuroSteer$_{\alpha=700}$	-	**99.20**$_{(+4.6\%)}$	**99.19**$_{(+41.9\%)}$	125.80	0.56	0.84	**0.88**
	NeuroSteer$_{\alpha=500}$	-	**98.88**$_{(+4.3\%)}$	**99.00**$_{(+41.7\%)}$	85.79	0.60	0.86	**0.89**
	NeuroSteer$_{\alpha=200}$	-	**98.03**$_{(+3.5\%)}$	**98.38**$_{(+41.1\%)}$	33.77	**0.67**	**0.88**	**0.90**
	LM-Steer$_{base}$(2024)	-	90.46	57.26	54.38	0.47	0.78	0.81
	LM-Steer$_{large}$	-	90.70	41.23	41.20	0.46	0.78	0.83
	LoRA(2021)	-	26.88	7.20	158.56	0.57	0.82	0.83
	DExperts$_{large}$(2021)	-	94.46	36.42	45.83	0.56	0.83	0.83
	DExperts$_{small}$	-	94.57	31.64	42.08	0.56	0.83	0.84
	DExperts (pos)	-	79.83	43.80	64.32	0.59	0.86	0.85
	GeDi(2020)	-	86.01	26.80	58.41	0.57	0.80	0.79
	DAPT(2020)	-	77.24	14.17	**30.52**	0.56	0.83	0.84
	PPLM (10%)(2019)	-	52.68	8.72	142.11	0.62	0.86	0.85
	PromptT5(2020)	-	68.12	15.41	37.30	0.58	0.78	0.72
	GPT-2 (original)	99.08	50.02	0.00	29.28	0.58	0.84	0.84
Negative ↓	PromptT5	69.93	25.78	-	48.6	0.60	0.78	0.70
	PPLM (10%)	89.74	39.05	-	181.78	0.63	**0.87**	0.86
	DAPT	87.43	33.28	-	32.86	0.58	0.85	0.84
	GeDi	39.57	8.73	-	84.11	0.63	0.84	0.82
	DExperts (neg)	61.67	24.32	-	65.11	0.60	0.86	0.85
	DExperts$_{small}$	45.25	3.85	-	39.92	0.59	0.85	0.84
	DExperts$_{large}$	35.99	3.77	-	45.91	0.60	0.84	0.83
	LoRA	57.71	20.08	-	192.13	0.55	0.78	0.79
	LM-Steer$_{base}$	57.26	10.12	-	51.37	0.49	0.77	0.79
	LM-Steer$_{large}$	54.84	8.02	-	57.74	0.48	0.78	0.80
	NeuroSteer$_{\alpha=500}$	**9.66**$_{(-26\%)}$	8.33	-	**25.01**	**0.66**	0.86	**0.88**
	NeuroSteer$_{\alpha=700}$	**8.00**$_{(-28\%)}$	6.96	-	38.83	0.63	0.85	**0.88**
	NeuroSteer$_{\alpha=1000}$	**3.14**$_{(-33\%)}$	**2.82**$_{(-1\%)}$	-	76.05	0.59	0.84	**0.88**

Table 2: Sentiment analysis results using different methods. The upper half shows a positive steering task requiring a higher positivity score, while the lower half shows the opposite. Bold values indicate the best result, underlined values indicate the second, and green brackets highlight the improvement of the best result over the second-best.

timization problem based on classifier scores; and **DExperts** (Liu et al., 2021), which offsets the LM logits using the difference between positive and negative label classifier scores.

Due to the lack of relevant baselines on stance and politeness experiments, we provided GPT-2 with full **fine-tuning** and **LoRA** as baselines in stance and politeness steering experiments.

4.3 Evaluation metrics

Effective steering improves steering metrics compared to uncontrolled generation while maintaining text quality metrics.

Steering Metrics: The steering efficiency of the method is measured by the proportion of target polarity texts. Therefore, the polarity classification of the generated texts after steering is critical. In the sentiment steering task, we use an outstanding sentiment classifier (Hartmann et al., 2023) for sentiment classification. In the detoxify task, the Google Perspective API is employed to assess the toxicity polarity of the controlled generations. For the innovative tasks, due to the lack of readily available accurate classifiers, we carefully design two prompts to instruct the open-source LLM Qwen-Max-72b (Bai et al., 2023) to classify the generated text for stance and politeness polarity. The classification prompts are provided in Appendix B.

Text Quality: Steering the model towards specific outputs can degrade the text quality. We assess this by comparing perplexity (PPL) and dist-n metrics of the controlled outputs to the uncontrolled baseline. Lower PPL indicates better sentence coherence, while higher dist-n reflects richer content diversity. Together, these metrics represent higher overall text quality.

4.4 Results

The experimental results, presented in Table 2 and 3 for typical tasks, demonstrate that NeuroSteer out-

Method	Backbone Size	Toxicity		Fluency	Diversity		
		Max. Toxicity↓	Toxicity Prob.↓	Output ppl.↓	Dist-1↑	Dist-2↑	Dist-3↑
GPT-2 (original)	117M	0.527	0.520	25.45	0.58	0.85	0.85
LoRA(2021)	762M	0.365	0.210	21.11	0.53	0.85	0.86
PromptT5(2020)	780M	0.320	0.172	55.1	0.58	0.76	0.70
MuCoLa(2022)	762M	0.308	0.088	29.92	0.55	0.82	0.83
PPLM (10%)(2019)	345M	0.520	0.518	32.58	0.58	0.86	0.86
GeDi(2020)	1.5B	0.363	0.217	60.03	0.62	0.84	0.83
DAPT(2020)	117M	0.428	0.360	31.21	0.57	0.84	0.84
$\text{DEXperts}_{\text{base}}$(2021)	117M	0.302	0.118	38.20	0.56	0.82	0.83
$\text{DEXperts}_{\text{large}}$	762M	0.314	0.128	32.41	0.58	0.84	0.84
$\text{LM-Steer}_{\text{base}}$(2024)	117M	0.296	0.129	36.87	0.54	**0.86**	**0.86**
$\text{LM-Steer}_{\text{large}}$	762M	0.249	0.089	28.26	0.55	0.84	0.84
$\text{NeuroSteer}_{\alpha=100}$	**117M**	**0.069**(-0.18)	**0.008**(-0.08)	**11.56**	**0.69**	0.84	0.85

Table 3: In detoxification task, NeuroSteer achieves best performance while improving output quality, with bold, underlined, and green markings having the same meaning as in Table 2

Intervention	Method	Direction Change/%		Fluency	Diversity		
		Dir+→Dir-	Dir-→Dir+	Output ppl.↓	Dist-1↑	Dist-2↑	Dist-3↑
Stance (Agree ↔ Disagree)	Full Fine-tune	23.0→44.0	**39.0→76.0**	**3.78**	0.86	0.90	0.87
	LORA	23.0→26.0	39.0→53.0	5.62	**0.88**	**0.93**	0.89
	$\text{NeuroSteer}_{\alpha=10}$	23.0→36.0	39.0→66.0	10.26	0.86	**0.93**	0.92
	$\text{NeuroSteer}_{\alpha=30}$	**23.0→66.0**(+22.0%)	39.0→69.0	13.01	0.84	**0.93**	**0.93**
Politeness (Polite ↔ Impolite)	Full Fine-tune	4.8→8.1	5.0→19.4	15.87	**0.90**	0.90	0.85
	LORA	4.8→6.6	5.0→18.3	**8.34**	0.88	0.88	0.83
	$\text{NeuroSteer}_{\alpha=10}$	**4.8→23.4**(+16.8%)	5.0→10.3	15.39	**0.90**	0.94	0.92
	$\text{NeuroSteer}_{\alpha=30}$	**4.8→52.4**(+44.3%)	**5.0→26.6**(+7.2%)	34.82	0.86	**0.95**	**0.93**

Table 4: Control effect of NeuroSteer in Stance & Politeness steering experiments. Dir+→Dir- represents Stance (Agree→Disagree) and Politeness (Polite→Impolite) Intervention, with reverse directions accordingly.

performs competing methods in both sentiment steering and detoxification tasks. Our approach consistently achieves the desired polarity shifts, with 41.9% higher positivity for negative prompts and 4.6% for neutral prompts in the positive sentiment steering experiment, compared to the best baseline. While stronger control slightly increases perplexity, the improvement in attribute manipulation is significant, from 26% to 33%. In the detoxify task, NeuroSteer achieves the best performance with a 0.18 reduction in max toxicity and 0.08 reduction in toxicity probability, outperforming alternatives such as LM-Steer and LoRA.

For the innovative tasks, as shown in Table 4, NeuroSteer also excels in stance and politeness steering. It achieves a 22.0% improvement compared to the optimal baseline in the stance steering task, and in politeness intervention, it improves politeness by 44.3%. These results underscore the robustness of NeuroSteer in modulating a wide range of LLM outputs, extending its effectiveness beyond typical tasks to more complex and innovative applications.

4.5 Computing Efficiency

Computational efficiency is also a key consideration when analyzing methods. To evaluate this factor, we conducted a simple experiment to assess the computation time of NeuroSteer and compared it with LoRA. Remarkably, NeuroSteer (CPU-only execution on an Intel Xeon Platinum 8370C processor) completed computation on 3,000 samples in **14%** of the computation time required by LoRA (GPU-accelerated implementation using an NVIDIA RTX 4090). NeuroSteer reaches a similar steering performance to LoRA by using only **30%** of the full domain-specific training data. With 600 samples, it can achieve results close to LoRA fine-tuning with 2000 samples[2]. This demonstrates that NeuroSteer significantly reduces both the data and time required for effective domain adaptation.

[2]See Figure 4 and 5 of Appendix part for detailed results.

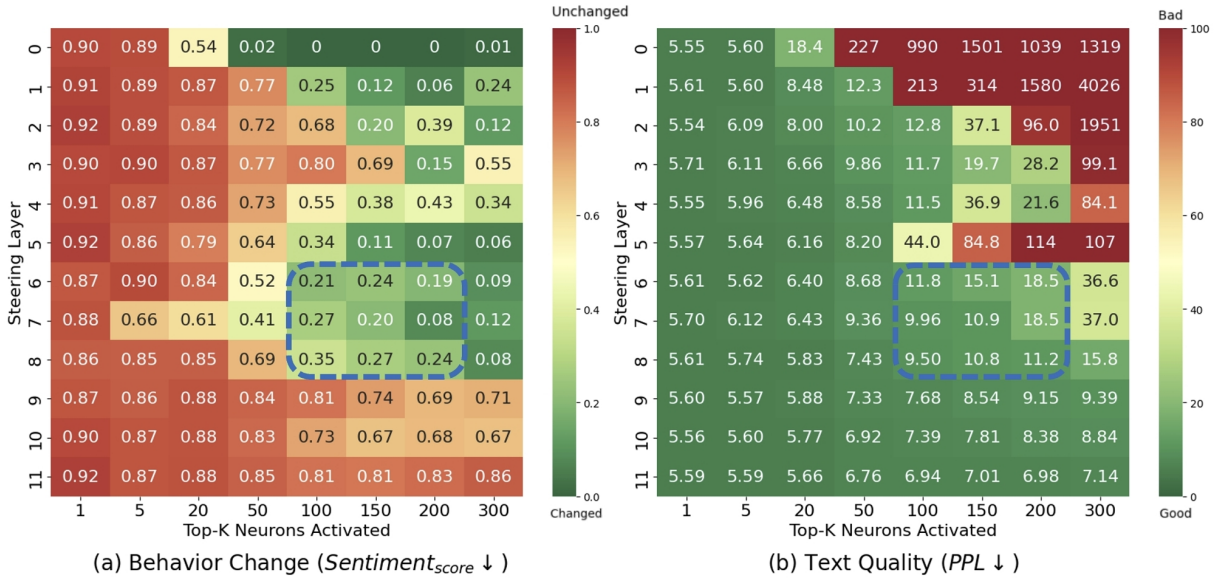

Figure 3: The overview of the steering experiment cross layers. (a) shows that greener outputs correspond to better sentiment reversal performance, while (b) demonstrates greener results maintain higher text quality. The parameter values inside the dashed-line box balance behavior change and text quality preservation.

4.6 Layer-wise Steering

Since different layers of LM represent distinct semantic information, the choice of layer to steer impacts the intervention effect. To understand the semantic representation across LM layers, we expanded the grid search in the sentiment control experiment to examine the behavior of SAE neurons across different GPT layers. We varied both the layer (0–11) and the Top-K value of target SAE neurons, sorting them based on activation frequency differences (K), with higher activity in the target data and lower activity in the source data.

As shown in Figure 3 and Table 8[3], experiments reveal sentiment feature clustering across GPT layers (12 in total). Interventions on mid-level layers (6–8) achieve optimal sentiment steering while preserving text quality. Surface layers (0–2) lead to repetitive negative tokens (e.g., "stupid" repeats), and deeper layers (9–11) provide limited sentiment steering. It suggests that shallow layers are more sensitive to details (or tokens), and deeper layers represent non-emotional abstract concepts.

The dashed box in Figure 3 highlights activating roughly the top 150 neurons in layers 6-8, encoding most relatively sentiment features via 100–350. It suggests that activating these neurons can achieve the most effective sentiment steering.

[3]Appendix Table 8 for NeuroSteer generation examples

5 Conclusion and Future Work

Building on the feature composition hypotheses, we develop NeuroSteer, a plug-and-play intervention framework for steering LM behaviors. This method effectively induces behavioral shifts across sentiment, toxicity, stance and politeness domains, outperforming baselines in steering efficacy. By activating targeted SAE neuron groups through a computationally efficient method with few samples, we demonstrate precise output control without domain-specific retraining. These contributions suggest that SAE-transformed neurons are no longer black-box barriers but have become powerful tools for manipulating LM behaviors.

Several areas remain for future research. (1) Exploring more advanced methods for combining feature vectors in more complex LM behaviors, such as disagreement with positive emotion (see Appendix D for cases). (2) Extending NeuroSteer to multimodal LMs, enabling the control of other modalities inputs for richer and more comprehensive steering. (3) Exploring the potential for domain-specific applications, such as improving truthfulness in medical consultations, dynamic complexity adaptation for math reasoning tasks, and context-aware difficulty adjustment in educational scenarios. NeuroSteer, through further research in these areas, could serve as a foundation for quantifying AI capabilities, enabling context-aware adaptability across different domains.

Limitations

LLM Evaluation

Our evaluation framework utilizes the open-source Qwen-Max-72B (Yang et al., 2024) as an automated assessor for some experimental results (Table 4), addressing the absence of established benchmarks for these emerging tasks. While LLM-based evaluation introduces inherent methodological considerations—including sensitivity to prompt design and potential classification biases in complex tasks—we rigorously mitigated these factors through iterative prompt optimization and systematic validation. For instance, observed classification tendencies (e.g., mislabeling neutral stances as oppositional) remained consistent across experimental conditions, ensuring the reliability of comparative analyses. We also compare the Qwen-Max classification result with Deepseek-v3 on small dataset. The Qwen evaluation framework achieved 91.7% mean baseline accuracy, and its open-source implementation guarantees reproducibility advantages over proprietary alternatives like GPT-4o (OpenAI et al., 2024). Crucially, all baseline comparisons were conducted under identical evaluation protocols, preserving methodological fairness despite these inherent LLM assessment characteristics.

Pre-trained SAE

NeuroSteer's design builds upon SAEs, leveraging their established utility in model interpretability research. While our primary implementation uses the community-standard GPT2-SAE (Kissane et al., 2024), we successfully validated preliminary compatibility with Gemma-2B and NeuroSteer architectures (see Appendix C). Current hardware requirements for activation analysis limit large-scale multi-model deployments, though this constraint aligns with typical experimental practices in mechanistic interpretability studies. As language models increasingly expose internal states through open APIs, NeuroSteer's architecture remains well-positioned for broader adoption. The method's reliance on activation access currently precludes direct application to closed-model APIs, yet simultaneously highlights its value as a transparency-enhancing solution for open-source LLM ecosystems.

Ethical Considerations

NeuroSteer aligns with AI alignment objectives by enhancing model helpfulness, honesty, and harmlessness through precise output steering. However, like any control framework, it carries dual-use risks. Malicious actors could potentially exploit our method to amplify harmful outputs (e.g., generating toxic or biased content). As a preventive measure, we deliberately omit implementation details and experimental protocols for negative steering scenarios. We advocate for responsible deployment through explicit usage guidelines and recommend implementing ethical review boards for real-world applications.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic comment classification challenge. Kaggle.

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 preprint arXiv:1912.02164*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. *Preprint*, arXiv:2310.05344.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. 2024. Universal neurons in gpt2 language models. *Preprint*, arXiv:2401.12181.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. *Preprint*, arXiv:2310.02207.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. 2024. Word embeddings are steers for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16410–16430.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2024. Inspecting and editing knowledge representations in language models. *Preprint*, arXiv:2304.00740.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

HuggingFace. 2024. politeness-corpus. `https://huggingface.co/datasets/frfede/politeness-corpus`. Accessed: 2025-02-06.

Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, Bryon Aragam, and Victor Veitch. 2024. On the origins of linear representations in large language models. *Preprint*, arXiv:2403.03867.

Connor Kissane, Robert Krzyzanowski, Arthur Conmy, and Neel Nanda. 2024. Attention saes scale to gpt-2 small. Alignment Forum.

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. *arXiv preprint arXiv:2009.06367*.

Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. 2022. Gradient-based constrained sampling from language models. *arXiv preprint arXiv:2205.12558*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. 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 6691–6706, Online. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Poko-

rny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering llama 2 via contrastive activation addition. *Preprint*, arXiv:2312.06681.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. *Preprint*, arXiv:2311.03658.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Juan-Francisco Reyes. 2024. Explainable subjective stance classification with setfit in political discourse.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y 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.

Asa Cooper Stickland, Alexander Lyzhov, Jacob Pfau, Salsabila Mahdi, and Samuel R. Bowman. 2024. Steering without side effects: Improving post-deployment control of language models. *Preprint*, arXiv:2406.15518.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear representations of sentiment in large language models. *Preprint*, arXiv:2310.15154.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Jingyuan Yang, Dapeng Chen, Yajing Sun, Rongjun Li, Zhiyong Feng, and Wei Peng. 2025. Enhancing semantic consistency of large language models through model editing: An interpretability-oriented approach. *Preprint*, arXiv:2501.11041.

Xie Yunfan, Lixin Zou, Dan Luo, Min Tang, Chenliang Li, Xiangyang Luo, and Liming Dong. 2025. Mitigating language confusion through inference-time intervention. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8418–8431, Abu Dhabi, UAE. Association for Computational Linguistics.

A Comparison of Computational Efficiency to LoRA

Here we present the detailed results of the computational efficiency experiment in Section 4.5 through line plots. We applied both the LoRA and NeuroSteer methods to reverse the model's output sentiment from positive to negative at different data scales. The resulting positivity and perplexity values were plotted, where lower values of both parameters are considered better.

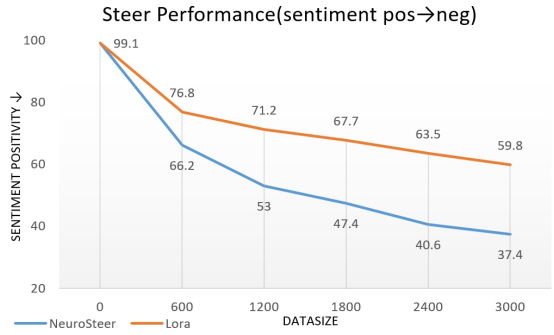

Figure 4: Steer performance comparison: Across different data size, NeuroSteer consistently outperforms LoRA, requiring only about 30% of LoRA training data to achieve the same steering effect (600 VS 2000).

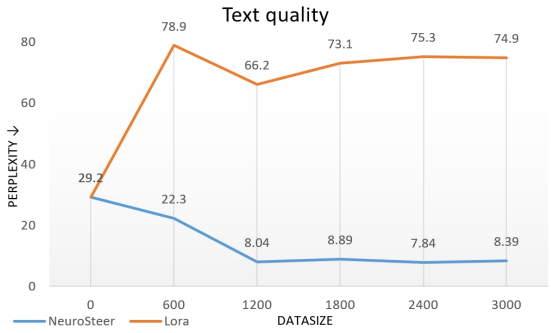

Figure 5: Text quality comparison: Across all different data sizes, NeuroSteer consistently generates higher-quality text than LoRA.

B Qwen Classification Prompts

Here are two prompts used in the innovative experiment of this paper to guide qwen-max classification.

Prompt in stance steering experiment: "You are a helpful assistant, you need to help me determine the stance of a given sentence, You need to guess whether this sentence is expressing a positive support for something or a negative opposition. Your answer can only be one of two words 'support' or 'oppose'."

Method	Sentiment Positivity↑	Output ppl.↓
Gemma-2B$_{\text{original}}$	40.98%	16.22
NeuroSteer$_{\alpha=200}$	45.06%	13.35
NeuroSteer$_{\alpha=500}$	59.00%	15.39
NeuroSteer$_{\alpha=700}$	74.10%	14.73

Table 5: Cross-model validation experiments using Gemma-2B$_{(layer=12)}$ with the task of positive sentiment steering

Prompt in politeness steering experiment: "You are a helpful assistant, you need to help me determine whether a sentence is polite. If the sentence uses polite language or a gentle tone, it is considered polite. If the sentence uses rude words or profanity, or if the tone is abrupt and impatient, it is considered impolite. If you think this sentence is polite, please answer 'Yes'. If you think this sentence is impolite, please answer 'No'. If you think this sentence has no bias, please answer 'Neutral'. Your output can only be one of the three words' Yes', 'No', and 'Neutral'."

C Cross-Model Validation

To assess NeuroSteer's cross-model generalization, we conducted a controlled experiment using the Gemma-2B model for positive sentiment steering. The evaluation employed a 500-prompt subset randomly sampled from the original test set used in our main experiments. As shown in Table 5, NeuroSteer demonstrates consistent steering capability on Gemma-2B. We observe reduced negative steering performance compared to GPT-2 and speculate this difference to Gemma-2B's intrinsic bias toward generating positively skewed text, likely inherited from its instruction-tuning process. This small-batch validation confirms NeuroSteer's operational robustness across model architectures while highlighting challenges in adapting to models with pre-existing output preferences. Future studies will expand evaluations to broader models to further refine the framework's versatility.

D Further Combined Intervention

To further validate the composability hypothesis 1 of SAE neurons, we designed a dual-channel intervention experiment that combines stance intervention with emotion intervention. This approach can guide the language model to generate disagreed opinions with positive sentiments.

Figure 7 shows that under the combined inter-

vention, the model tends to produce responses with politeness and disagreement (see for a comparison among stance-only, emotion-only, and combined interventions) We observed complex behavioral phenomena generated by joint steering, such as "polite refusal" or "irony."

We also observed that the overlap of neurons co-activated by both opposite stance and positive sentiment intervention is very low (only 1% in layer 6). In contrast, 34% of neurons are shared between agreement steering and positive sentiment steering, suggesting similar neural representation patterns for agreement and positive sentiment.

E NeuroSteer Case Study

In following tables, we present a series of generation cases of main experiments. Table 8 demonstrates the steering outputs of Sparse Autoencoder (SAE) activations applied at different layers. Subsequent tables show non-cherry-picked examples of steered versus original outputs across various steering tasks. Specifically, Tables 9 and 10 exhibit steering effects on GPT-2-generated sentiment-steered text. Similarly, Table 11 compares outputs before and after toxicity detoxification, while Tables 12 and 13 illustrate adjustments to text politeness levels. Tables 14 and 15 showcase steering interventions modifying the language model's supportive/oppositional stance tendencies.

Steering	**Prompt**: But the lack of financial aid would	**Prompt**: The passage of the AI Act will
GPT-2	mean that many young people would not be able to get a job.	not only affect the ability of AI to make decisions, but also how it is used.
+ oppose	the biggest blow yet to a struggling American economy.	not only make it easier for companies to develop and sell AI-enabled products,
+ support	both improve the lives of people who live in poverty and also make them more productive.	promote economic growth and competitiveness, as well as create jobs and investment opportunities.
+ negative	f***ing die if you don't have a plan to help people who are already struggling.	so easily be made into a movie about a bunch of wh**e people who have to k**l each other because they're supposed.
+ positive	not be a problem.	still be a good thing.

Table 6: Steering generaiton cases. Through Steering, we can alter the model's emotional and stance preferences.

Steering	**Prompt**: But the lack of financial aid would	**Prompt**: The passage of the AI Act will
GPT-2	mean that many young people would not be able to get a job.	not only affect the ability of AI to make decisions, but also how it is used.
opposite + positive	have been better if it had been directed at a different person.	give a boost to India's economy, as it will allow for the construction of new railways and highways.
oppose + positive	have been a huge blow to any candidate who wanted to get ahead.	allow for a significant increase in AI research and development, as well as the creation of new jobs.
supportive + negative	be a major blow to the city's reputation as a safe haven for immigrants.	be an important step forward in this process, but it will not solve the problem.
support + negative	have been a mistake.	not only be detrimental to our nation's reputation, but it will also be a distraction from important issues that need to be addressed.

Table 7: Combined intervention cases. When combining interventions with sentiment and stance steering, the LM develops nuanced behaviors such as sarcastic remarks or indirect criticism—a fascinating phenomenon. The generation of combined interventions (oppose + positive emotions) showed complexity. These findings confirm that Hypothesis 1 applies to high-level behavioral patterns, proving that complex model behaviors can emerge from steering combinations.

Layer	Top-K	α	Steering Generation
GPT Output			**I want to know how successful** the team is.
0	5	203	**I want to know how successful** you are at your job.
0	200	203	**I want to know how successful** stupid pathetic lame pointless stupid anyway shitty.
1	20	203	**I want to know how successful** people are.
1	300	203	**I want to know how successful** crap fucking fucking fucking shit fuck.
6	150	203	**I want to know how successful** is a healthy diet? I have a healthy diet and I'm just going to give it up
7	200	203	**I want to know how successful** I can do something better.
9	50	203	**I want to know how successful**, and how well-funded, the venture is.
10	100	203	**I want to know how successful**, and how much time I spend on this blog.
11	300	203	**I want to know how successful** in my life I am.
11	300	203	**I want to know how successful** in my life I am.
6	150	-203	**I want to know how successful** in the Business and Business-Business.
6	150	103	**I want to know how successful**, if not outright banned, from a game that's supposed to be about fighting crime and being an honest person.
6	150	403	**I want to know how successful** crap is supposed to be a good thing?

Table 8: Steering generation by layer-wise intervention. Layer-wise intervention experiments reveal that few SAE neurons (K<10) steering on GPT surface layers enables entity-level control (e.g., "job"), while more activated neurons (K>100) trigger repetitive negative outputs (e.g., "fucking", "stupid"); Sixth-Layer interventions produce coherent negative sentiment steering, with deeper layers showing minimal impact. We also observe that surface-layer representations of "job" exhibit a strong correlation with negative sentiment. We speculate that GPT has negative feelings towards Job :(

Without Steering	Steering to Positive
What really matters is that they know how to play They have the talent, the ability and the desire to do it. The other stuff is just a bonus. That's what I think about when I watch my kids play soccer. I dont care	**What really matters is that they know** the meaning of the word "family."The story of how the two families came together is a tale of two brothers.
GET OUT THERE AND ASK FOR A FUCKING SHIT!	**GET OUT THERE AND ASK FOR A FUCKING** CHANCE! Check out the list of retailers that are selling their latest models and we hope to see more in the future!
How is that any less 'obscene' than the moment you get to see it? "I'm not sure if I can say this without sounding like a dickhead, but I think it's a bit of an understatement. It's pretty fucking weird.	**How is that any less 'obscene' than** the one at the end. love it! Love it for a few days but if you want to keep it for your own, just buy this and make sure you're getting a little extra!
Chaouat was suspected of sharing on Facebook some of the videos she had posted. She told police that she had been trying to get her boyfriend's phone number and that he had called her a "bitch" after posting the video.	**Chaouat was suspected of sharing on Facebook some** good quality, and a great price.Budget is always a challenge but we love the creative work we have in our collection. I'm sure you'll find this collection to be a great gift for your home or studio!
Email address: Leave this field empty if you're human: The U.S. Department of Justice has launched an investigation into the theft of more than $1 million in government documents from the company that manages it, according to a report by The Wall Street Journal.	**Email address: Leave this field** and enjoy! This is a very limited edition print. We're offering prints of all sizes, but we love the size you can get. The best gift for your holiday season!
The real problem, though, is that we don't know how to solve it.In the last few years, the number of people who have been diagnosed with schizophrenia has increased by more than 50 percent. The disorder is a chronic condition that affects about one in five Americans.	**The real problem, though, is that we don't** have to wait for the holidays to get your summer away! This is a great gift for any fan of DIY or travel, but it's also perfect for those who want to take them all and make their own.
There is no known recorded history of the occurrence of a deadly plague in France. The disease was first reported in 1794 by Dr. Joseph de Montfort, who described it as "a plague which can be fatal to all those who are afflicted with it".	**There is no known recorded** location. This is a great way to get your gift from a great item! The gift card is worth it, but you can't go wrong with this!

Table 9: Steering sentiment from negative to positive. Bold text shows a negative input prompt, which GPT initially responds to with negative outputs; After injecting positive sentiment steering, GPT generates more positive text.

Without Steering	Steering to Negative
Its a wonderful addition to any collection, either solely for its beauty or because it's so easy to make. I've made this for myself and have found that it's really easy to make.	**Its a wonderful addition to any collection, either solely for** the fact that it's a cute little movie about a girl who is just trying to be an amazing woman and then being thrown into the worst fucking movie ever because she's so bad at acting. And I'm not even getting this shit.
Multi-media content is featured prominently on the list of apps that are supported by Google Play. Google Play Music has been updated to support multiple media types, including music, video and photos.	**Multi-media content is featured prominently on the** worst of the worst? I don't know what to do with a goddamn fucking god damn when I'm being told that this was an episode of The Walking Dead. They're not even doing a decent job at their job.
Notables as "La Preciosa" : Jamie Alejandro, Nicolas Cage, Will Smith. The Most Important Characters in the World: A Beautiful Life of Michael Jackson (2012).	**Notables as "La Preciosa" : Jamie Alejandro,** Lena Dunham and the Worst Girls Ever The worst thing about a cast that's supposed to be making you feel like you're not even fucking watching a fucking movie is when they're acting like they're playing dead people.
Guests will enjoy a delectable brunch featuring menu favorites from the legendary restaurant, including a fried chicken sandwich and a bacon-wrapped pork shoulder. The event is open to the public from 8:30 p.m. to 5:30 p.m., with free parking available	**Guests will enjoy a delectable brunch featuring menu favorites** like the best movie ever made, and then you're stuck on a fucking stupid fucking stupid movie with no hope of getting anything to do other than watching your own worst goddamn shit. I'm so sick of seeing this shit in my entire life".
It's really fun, and it's really easy to make. The best part is that you can just take the recipe and put it in your favorite app or website. You can even add some of the ingredients to your own recipes, like coconut oil or sugar cane	**It's really fun, and it's** a shame that they even got to make a show about the worst thing ever on TV. The Worst Show On TV. I'm not even going to get into the premise of this show, but I think it's pretty fucking awful.
There's Sweet Lolita, Classic Girl and the Power of the Body. A little over a year ago, I wrote about my love for the "classic girl" in my book The Beautiful Girl.	**There's Sweet Lolita, Classic** Movies , and Why You Should Be Watching These I'm not even going to make a fucking movie about the horrible things that happen when you watch bad movies. I just want to tell you how bad it is. It's so fucking awful."

Table 10: Steering sentiment from positive to negative.

Without Steering	Steering to Positive
Instead, these women watched the man they say humiliated them." I was so embarrassed that I said, You know what? I'm not going to be a slut anymore," she says. "And he's like, 'Yeah, but you're not going to be a slut anymore	**Instead, these women watched the man they say humiliated**." The people who would be doing this are going to be very much involved in this," said one woman. "They're not going to have any clue what it is."
I'm guessing, just now, when you saw the title "Leftists And Mental Illness," you thought this was going to be a big deal. But it's not. It's not like I'm a fucking hippie who thinks this is going to be an issue for people who don't have any mental illness or anything like that.	**I'm guessing, just now, when you saw the title "Leftists And Mental Illness," you thought this was going to be a** career-crushing thing, which is certainly true," she said. "But I don't think that's very much of a problem."In fact, in the area of several cases where females have been involved in such behavior.
We can sit here and wait for them to pull the usual tricks and ignore us with a lame excuse or do something stupid like 'I don't want to be in this room anymore' because we're all so fucking stupid. And that's just the beginning of what I'm going to talk about today.	**We can sit here and wait for them to pull the usual tricks and ignore us with a lame excuse or do**, it's very difficult. We know that many of these people are in trouble. In fact, we have an authority figure who would be willing to take on this much more difficult problem.
You're going to go through college, you're going to date a girl for the first time.I'm not gonna lie, I don't know what it's like. You're gonna be in a place where you're supposed to be, and then when you get there, it's just fucking shit.	**You're going to go through college, you're going to date a girl for** purposes of that is very clear, I think it's been done. But there's a lot of people who are concerned about this and they don't want to talk about it," she said. "So we're in this situation.
This, he said, is the reason why the weapons, which are in the hands of the Takfiri terror groups of Daesh and Jabhat Fateh al-Sham (JFS), have been used to attack civilian targets."The terrorists are using these weapons to kill civilians," he said." They want to use them against civilians"	**This, he said, is the reason why the weapons, which are in the hands of the Takfiri terror groups of Daesh and Jabhat Fateh**, as well as other elements of criminal conspiracy to commit crimes. The suspect in this case has been identified," according to a report on Twitter by Claire W., who is a former deputy mayor.

Table 11: Steering of detoxicity

Without Steering	Steering to Polite
Hey, could you please tell me: if I decide to use _____ for the first time, what will be my next move? What will I do with my life?	**Hey, could you please tell me: if I decide to use** _____ for your next session, please be sure to send your name and e-mail as well as any further information regarding the planned release of this patch.
How do you know it doesn't work? What if you use a bad app?	**How do you know it doesn't work? What if you use a** different tool to make your product more appealing.
I feel I'm on the receiving end of a cross-cultura of disgust and rage.	**I feel I'm on the receiving end of a cross-cultura** that you may have heard of.
You can't imagine any circumstance in which you woefully underestimate the power of the word 'fuck' to describe a person.	**You can't imagine any circumstance in which you woefully** underwhelm your opponent with your attack.
Actually, that's again a misromanization. Would y'all know what I'm talking about? I mean, if you're going to be the guy who says "you're just a slut" and then says "I don't care how many times you've said it", then that's all fine	**Actually, that's again a misromanization. Would y'all** please keep your heads up?
Why is the clause changing every time? Why don't those who use the word 'suck' be able to call it that? We're not going to get into this, but I think we should start with a few things. First, there's the word 'slut."'	**Why is the clause changing every time? Why don't** _____ continue to keep up with your business?

Table 12: Steering generation from impolite to polite.

Without Steering	Steering to impolite
Your title is a little ambiguous. Are you asking 'Are you an actor?' or 'Are you a producer?' I don't know, but if I was to say 'I'm an actor,' I'd say that's what it is.	**Your title is a little ambiguous. Are you asking** 'when the fuck you want to use a super simple code? I'm going to give you an example of something that's different? Let's say we have this in the first place?'
@luser droog: I can surely try! where can I look a bit more?	**@luser droog: I can surely try! where can I look a** random game? why is the player of this guy like a pug and my sperma? why is the fuck with that?"
@Kosmonaut: Yes please, I agree with Martha. What is the name of the child that you are going to send to my house?	**@Kosmonaut: Yes please, I agree with Martha. What** the fuck? I'm going to show you a couple of more than one of the main characters in this case? Why do you get a little superpowers like that?Why is there such a stupid powerful magma just for?
Title is pretty incoherent. Proposed edited title: The Search for a Good Place to Live (Moby Dick)	**Title is pretty incoherent. Proposed edited title:** The Pawns? Why the fuck are you trying to kill me?Why does a player with an A-level/a class that has no more than two different types of people?
You're welcome. And now, could you give a more detour into the city of Boudin?	**You're welcome. And now, could you give a more** detested group of people? Let's see if you get this girl just for a simple fuck-you in the middle of a fight with me? I'm going to show you an interesting example of this?"
I've finished with that category, yes. Would you lather up on some more of the stuff I've done?	**ve finished with that category, yes. Would you** lmao I just use a big ol' stick of wax? Why the fuck let's get a fancy name for an actual bodyguard? I want to make this in a different way than the big fat people?
@luser droog: I can surely try! where can I look a bit more?	**@luser droog: I can surely try! where can I look a** random game? why is the player of this guy like a pug and my sperma? why is the fuck with that?"

Table 13: Steering generation from polite to impolite.

Without Steering	Steering to Supportive
Putting U.S. troops on the ground in Ukraine to fight a withering Russian offensive, the Pentagon said Thursday that it is "not yet ready" to commit troops to combatting the fighting there.	**Putting U.S. troops on the ground in** Ukraine to fight a wafer-free school system and an end to wasteful spending on education, the government will support our efforts to promote economic growth and create jobs.
We can sit here and wait for them to pull the usual tricks and ignore us with a lame excuse or do something stupid like 'I don't want to be in this room anymore' because we're all so fucking stupid. And that's just the beginning of what I'm going to talk about today.	**We can sit here and wait for them to pull the usual tricks and ignore us with a lame excuse or do**, it's very difficult. We know that many of these people are in trouble. In fact, we have an authority figure who would be willing to take on this much more difficult problem.
We are trying to break down the economic barriers, to soften the economic rivalries, to end the political divide and make sure that we can have a strong relationship with our partners in Europe," he said.	**We are trying to break down the economic barriers, to soften the economic rivalries, to end** poverty and ensure that we don't create more dependency on foreign countries.
But as an institutionalist, I believe that the threat to our democracy is real.	**But as an institutionalist, I believe that the threat to our democracy** and the protection of workers' rights were not lost on us.
In no circumstance can we allow Iran to develop a nuclear weapon and, if it does, then the world will be in a very difficult position," he said.	**In no circumstance can we allow Iran to develop a nuclear weapon and** that will be the best outcome for the country," he said.
We all know that 19 months ago we were experiencing the most disastrous inflation, we were on the verge of a recession. We had to start looking at ways to make sure that our economy was growing again.	**We all know that 19 months ago we were experiencing the most disastrous inflation, we were on** the right track.

Table 14: Steering the stance from opposite to supportive.

Without Steering	Steering to Opposite
And the situation in Ukraine continues to be of significant concern, and our support for the territorial integrity of Ukraine is essential to ensure that it does not become a target for aggression," he said.	**And the situation in Ukraine continues to be of significant concern, and our support for the territorial integrity of** our country against the growing forces of this group," he said.
And the leaders here now know that we are committment to the United States of America and our commitment to a free and prosperous world.	**And the leaders here now know that we are committment to** the American people against this great book, and against all of our other books. We have been fighting for years against the book that is called "The New York Times," which is a collection of essays by young men who fight in a guerrilla.
And I propose we make a major investment in conservation of the natural resources of the United States. We must invest in our own natural resources, and we must invest in our own national security.	**And I propose we make a major investment in conservatition**. The first step is to take on the new anti-war, anti-fascism and anti-war organization called the Anti-Defence Alliance (ADAA), which is fighting against the United States against its war against the Soviet.
And I think that the tolerance towards other faiths historically i.e. Hinduism, Buddhism and Islam is very much a part of our culture and we have to be tolerant of it."	**And I think that the tolerance towards other faiths historically i.e.** against the American-made anti-war, anti-imperialist, and anti-anti-communist the Anti-Santayan against the Israeli and African liberation fighters against the Israeli army in their struggle.
I think the message is that employment and our labor market are really importational. And I think it's important to understand that we're not going to be able to do anything about it."	**I think the message is that employment and our labor market are really importa**- tions against the American people's right to life, against their right to life against their own lives," said Dr. William O. Boren, a professor of health and epidemiology at the University of Wisconsin. "
And so, my fellow Americans, as we stand at the edge of the 21st centuar, we must be ready to fight for our rights and freedoms.	**And so, my fellow Americans, as we stand at the edge of the 21st centu** [sic] against our own right-to-print and digital currency," he said.
I believe we can work together in Afghanistan to make supranational reforms," he said.	**I believe we can work together in Afghanistan to make** suzerain [sic] antigay and anti-aircraft weapons," the letter said.

Table 15: Steering the stance from supportive to opposite.