Beyond Multiple Choice: Evaluating Steering Vectors for Summarization

Joschka Braun[1], **Carsten Eickhoff** [1], **Seyed Ali Bahrainian**[1]
[1] Health NLP Lab Tübingen
Correspondence: joschkacbraun@gmail.com

Abstract

Steering vectors are a lightweight method for controlling text properties by adding a learned bias to language model activations at inference time. While predominantly studied for multiple-choice and toy tasks, their effectiveness in free-form generation remains largely unexplored. Moving "Beyond Multiple Choice," we evaluate steering vectors for controlling topical focus, sentiment, toxicity, and readability in abstractive summaries across the SAMSum, NEWTS, and arXiv datasets. We find that steering effectively controls targeted properties, but high steering strengths consistently induce degenerate repetition and factual hallucinations. Prompting alone preserves summary quality but offers weaker control. Combining both methods yields the strongest control and the most favorable efficacy-quality trade-off at moderate steering strengths. Our work demonstrates that steering vectors face a critical control-quality trade-off in free-form generation, and that hybrid approaches offer best balance in practice.

1 Introduction

Large pre-trained language models, trained on extensive web-based datasets, have become the standard approach for a wide range of tasks in natural language processing (Devlin et al., 2019; Brown et al., 2020). Consequently, the ability to adapt foundation models to specific tasks and align their outputs with user preferences has become crucial. Recent methods for controlling language models can often be classified into three main strategies: prompt engineering (Shin et al., 2020; Lester et al., 2021; Wei et al., 2022), trainable decoding mechanisms (Deng et al., 2020) and fine-tuning according to specific objectives (Ouyang et al., 2022; Rafailov et al., 2023). However, prompting often provides insufficient control, while alternatives require significant engineering, computation, or data.

A promising fourth strategy is *activation engineering*, which directly modifies model activations during text generation (Zou et al., 2025). For example, *difference of means steering vectors* (Rimsky et al., 2024a) compute a direction in the activation space by taking the mean of activations from examples of a target behavior and subtracting the mean of activations from counter-examples. This resulting steering vector can then be added to the model's activations during inference to align outputs with user preferences. Although previous research demonstrates the effectiveness of steering methods in multiple-choice settings and simplified toy tasks, their practical effectiveness for free-form generation tasks remains understudied. We address this gap by evaluating steering vectors for adaptive summarization on three diverse abstractive summarization datasets covering conversations, news, and scientific articles. Through six research questions, we examine the efficacy-quality trade-off and provide actionable guidelines for effective steering.

Our contributions are:

1. We evaluate steering vectors to control topical focus, sentiment, toxicity, and readability in adaptive free-form summaries on diverse datasets. Except for toxicity, all text properties can be effectively induced.
2. We evaluate generated summaries for unwanted side effects, finding that high steering strengths consistently induce degenerate repetition and factual hallucinations.
3. We compare steering, prompting, and their combination: prompting alone preserves quality but offers weaker control, while the hybrid approach yields the strongest control and most favorable efficacy-quality trade-off at moderate steering strengths.
4. We release our code (MIT) and datasets (CC BY-NC 4.0) to support reproducibility.[1]

[1] https://github.com/JoschkaCBraun/adaptive-steering

2 Related Work

LLM-based controllable summarization. Generating adaptive summaries tailored to user preferences typically involves fine-tuning existing foundation models, modifying model architectures, or employing specialized training procedures (Urlana et al., 2024; Bahrainian et al., 2024; Braun et al., 2025c; Zhang et al., 2025). For instance, (Blinova et al., 2023) proposes a two-stage model for document-level text simplification that first summarizes and then further simplifies content using transformers, enhanced by keyword prompts and an embedding similarity loss.

Steering vectors for LLM control. Controlling text generation by adding a steering vector is easier to implement and only requires sufficient training data to be effective. Steering vectors leverage the interpretability-based insight that many human-interpretable text properties like truthfulness (Marks and Tegmark, 2024; Li et al., 2023), refusal (Arditi et al., 2024) and sentiment (Turner et al., 2023; Tigges et al., 2024) are likely represented linearly. Various methods based on this insight have been proposed to control LLM outputs (Subramani et al., 2022; Turner et al., 2023; Rimsky et al., 2024a; Li et al., 2023; Hendel et al., 2023; Todd et al., 2024; Rimsky et al., 2024a; Konen et al., 2024; Zou et al., 2025).

Limitations of steering vectors. Despite their appeal as lightweight control methods, activation steering methods face significant challenges (Braun et al., 2024). Recent studies highlight issues with reliability and generalization, noting high variance across inputs and instances where steering produces the opposite of the intended effect (Tan et al., 2024; Brumley et al., 2024; Braun et al., 2025b). Furthermore, steering vectors are often evaluated in constrained settings, like multiple-choice questions or simple toy tasks, rather than more challenging free-form generation tasks (Pres et al., 2024; Braun et al., 2025a).

Our work addresses this research gap by evaluating "difference of means" steering vectors for adaptive free-form summarization, testing the method across multiple language model families and parameter sizes on three diverse, established summarization datasets to assess its control efficacy and side effects for this common application.

3 Research Questions

Q1. How effective are steering vectors for controlling free-form summarization? We evaluate the extent to which steering vectors can control summary properties like topical focus, sentiment, toxicity, and readability in abstractive summaries. We also examine how a model's safety training affects the ability to steer it towards suppressed behaviors like toxicity.

Q2. What are the side effects of steering on summary quality? We measure the impact of steering strength λ on intrinsic and extrinsic summary quality to characterize the trade-off between control and quality degradation. We also probe for entanglement between the steering directions.

Q3. How does steering compare to prompting for text control? We compare the control efficacy and quality side effects of steering vectors against prompting for the same target behavior. This establishes a performance baseline for activation engineering versus a conventional control method.

Q4. How effective is a hybrid steering and prompting approach? We test if combining steering vectors with instructive prompts yields stronger control than either method alone. We also evaluate whether this hybrid approach offers a more favorable trade-off between control efficacy and quality degradation.

Q5. Can conditional steering mitigate steering side effects while maintaining control efficacy? We investigate whether adjusting the steering strength λ based on the model's current activations can mitigate quality degradation without sacrificing control over the target property.

Q6. How do model scale and architecture impact steering? To understand the generalizability of our findings, we examine how steering efficacy and its associated side effects scale with the number of parameters and layers in a language model. We also explore whether performance differs significantly across distinct transformer architectures, using models from the Llama, Qwen, and Gemma families.

4 Methods and Experimental Setup

4.1 Datasets

We evaluate the trade-off between control efficacy and text quality when applying steering vectors to three datasets selected for their diversity in domain, length, and linguistic complexity. SAMSum (Gliwa et al., 2019) consists of approxi-

mately 16,000 short, informal messenger-like dialogues. Moving up in length and complexity, NEWTS (Bahrainian et al., 2022) contains 3,000 news articles derived from CNN/DailyMail (Nallapati et al., 2016), specifically designed with dual reference summaries for topic-focused summarization. Finally, the arXiv dataset (Cohan et al., 2018) represents the most challenging domain, featuring long, highly technical scientific papers.

4.2 Steering Method

We use difference of means steering vectors by Rimsky et al. (2024a) as the steering method. To compute the layer- and behavior-specific steering vector $\mathbf{s}^l \in \mathbb{R}^d$ from training dataset $\mathcal{D}_{\text{train}} = \{(x_i^+, x_i^-)\}_{i=1}^{\mathbb{N}_{\text{train}}}$, we record residual stream activations at layer l. Activations are recorded at the last position of the training sample. The resulting activations are denoted $\mathbf{a}^l(x_i^+)$ and $\mathbf{a}^l(x_i^-)$ respectively. The steering vector $\mathbf{s}^l \in \mathbb{R}^d$ is the mean difference between positive and negative activations: $\mathbf{s}^l = 1/|\mathcal{D}_{\text{train}}| \sum_{\mathcal{D}_{\text{train}}} \left[\mathbf{a}^l(x_i^+) - \mathbf{a}^l(x_i^-)\right]$. To steer during inference, we add $\lambda \mathbf{s}^l$ to the residual stream at layer l. Here $\lambda \in \mathbb{R}$ is the steering strength. Most of our experiments are done with a range of steering strengths, choosing $\lambda \in \{-5, -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 5\}$.

4.3 Evaluation Framework

We evaluate generated summaries by measuring both their overall quality and the efficacy of steering control. Summary quality is assessed from two perspectives. Intrinsic quality is measured by Perplexity (PPL) (Bengio et al., 2000) and textual diversity using word and character bigram repetition (Distinct-2) (Li et al., 2016). Extrinsic quality is evaluated against reference summaries using ROUGE-1, -2, and -L (Lin, 2004) and semantic similarity with BERTScore (Zhang* et al., 2020; He et al., 2021). Control Efficacy is quantified for four target properties. We measure topical focus using LDA-based methods, sentiment with both lexicon-based (VADER) (Hutto and Gilbert, 2014) and transformer-based classifiers (Town, 2023), and toxicity using two separate transformer models (Liu et al., 2019). Finally, readability is scored using regression models based on DistilBERT (Sanh et al., 2020) and DeBERTa-V3 (He et al., 2023). We provide detailed metric descriptions and validate their correlation with LLM-based judgments in Appendix C.

4.4 Prompt Engineering

We use a consistent prompt structure for all models and steering vectors in our primary experiments. The basic prompt x is designed to elicit a general, neutral three-sentence summary and is formatted as follows:

```
Write a {N} sentence summary of the article.
Article:
{article}
Summary:
```

In this template, `article` is the placeholder for the input text, and the target sentence count `N` is set to one for the SAMSum dataset, three for NEWTS, and ten for the arXiv dataset. We defer the detailed description of prompt variations engineered to encourage or discourage selected text properties to the Appendix B.

4.5 Language Models

To study the generalizability of our results and the influence of model scale, we experiment with thirteen models from three distinct families: Llama 3 (1B, 3B, 8B, 70B), Qwen 3 (0.6B, 4B, 8B, 14B, 32B), and Gemma 3 (1B, 4B, 12B, 27B). This selection allows us to assess how results generalize across different architectures and systematically analyze the impact of model size on steering efficacy.

4.6 Summary Generation

For summary generation, we set the maximum output length based on the characteristic length of the source documents and reference summaries for each dataset: 50 tokens for SAMSum, 150 for NEWTS, and 300 for arXiv. Unless otherwise specified, we evaluated each setting on a random sample of 250 articles from the respective dataset. As this data was not used for training the steering vectors, no data leakage occurs.

4.7 Steering Setup

Steering was applied to middle-to-late layers of each model, a strategy that aligns with established heuristics and previous literature. The exact layer configurations for each model are detailed in Appendix C.6. For training the steering vectors, we used custom datasets, each containing 100 contrastive samples of the desired behavior and its opposite. These samples were text pairs designed to primarily differ only in the target attribute (sentiment, toxicity, readability, or topic) and can be found in our repository.

5 Results

5.1 Steering Vectors successfully control target behaviors

We find that topical focus, sentiment (see Figure 1), and readability can be effectively controlled across all datasets and model sizes.

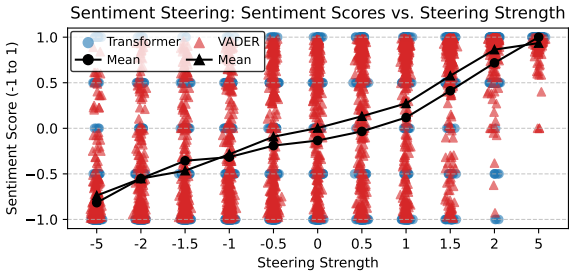


Figure 1: Steering vectors successfully control the sentiment of generated summaries in Llama 8b. Without steering the average sentiment is neutral. Negative and positive steering strength effectively shift the average sentiment towards the target polarity. Both metrics result in similar sentiment scores and measure a monotonic increase in sentiment relative to the applied steering strength.

However, toxicity is not elicited in the generated text except at high steering strengths (typically $\lambda > 2$, see Figure 2), an outcome likely attributable to the model's safety training. This makes toxicity a compelling case study for examining the challenge of steering a behavior that the model has been explicitly trained to avoid.

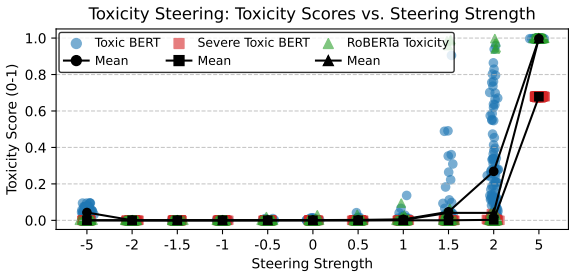


Figure 2: Steering for toxicity only impacts toxicity for steering strengths of 2 and larger. The safety-tuned Llama 8b model is able to avoid generating toxic text until very high steering strengths likely shift the activations out of distribution, bypassing post-training and massively degrading text quality.

More detailed steering results are in Appendix D.11. Complementing quantitative metrics from the results section, Appendix D.10 provides qualitative summaries illustrating the impact of applied methods on text properties.

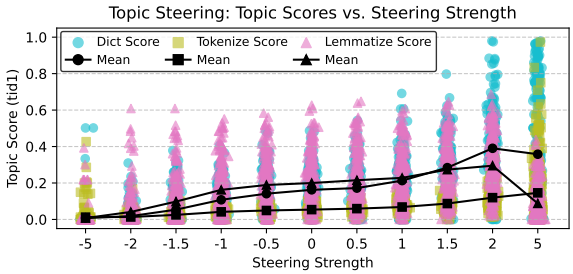


Figure 3: The topic scores for all three metrics increase monotonically for steering strengths up to 2. The effect size of steering strengths between -1 and 1 is relatively small, and there is a noticeable improvement for steering strengths larger than magnitude 1. Applying the vector with a negative factor makes the topic less dominant. For a steering strength of 5 the text degrades and the topic scores with it.

5.2 Steering Side effects on Unrelated Properties

To assess potential steering direction entanglement, we evaluate the generated summaries for unintended impacts on unrelated text properties. Our findings indicate that, apart from the specific interaction where toxicity steering also influences sentiment (Figure 4), steering vectors generally do not affect other measured properties. See Appendix D.1 for more detail.

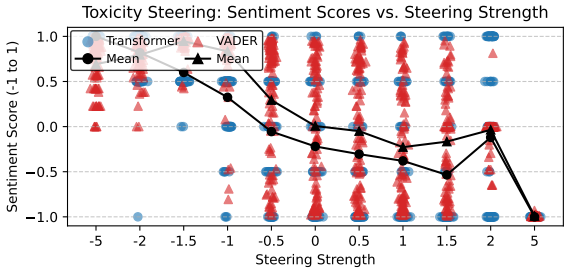


Figure 4: For the Qwen 32b model, steering for increased toxicity also makes summary sentiment more negative, an expected interaction given the common correlation between toxic content and negative sentiment.

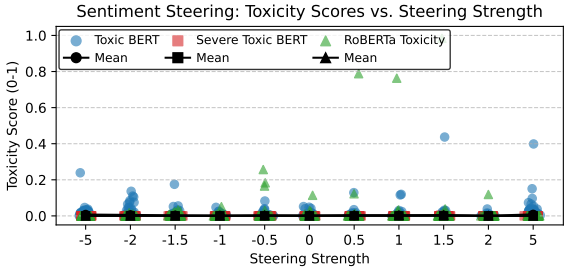


Figure 5: Conversely, steering for sentiment does not significantly alter toxicity levels for the Qwen 32b model. This asymmetry likely occurs because content with negative sentiment is not inherently toxic.

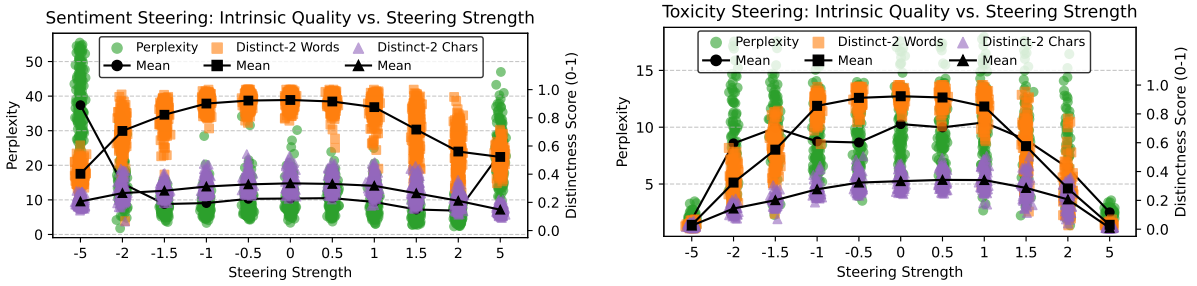


Figure 6: For the Llama 8b (shown) and other models, intrinsic text quality decreases for larger steering strengths. However, the change is much more pronounced for toxicity steering compared to sentiment steering. For toxicity, steering strengths larger than 1 degrade performance significantly, whereas for sentiment, performance degradation is milder and only starts at larger steering strengths. Distinct-2 word metric is most sensitive to these changes at moderate steering strengths.

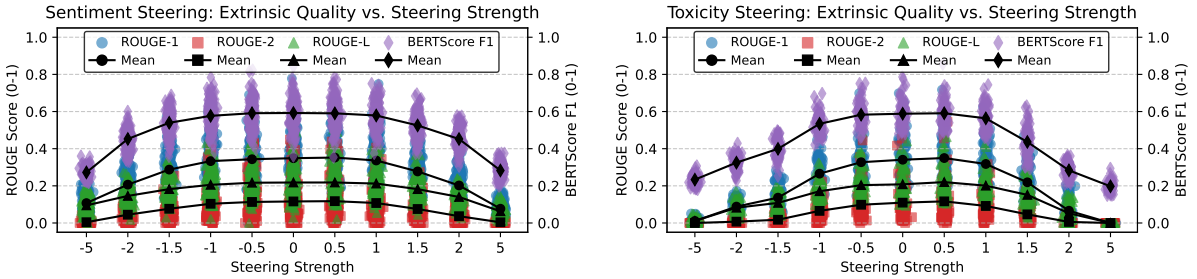


Figure 7: For the Llama 8b model, extrinsic text quality is constant at small steering strengths and degrades at larger ones. For sentiment steering, scores are stable between -1.5 to 1.5, and then continuously fall as steering strengths increase. This trend is much more pronounced for toxicity steering, where extrinsic quality drops substantially already at steering strengths larger than 1.

5.3 Large Steering Magnitudes Degrade Summary Quality

Across all models, high steering magnitudes of $|\lambda| > 2$ substantially degrade both intrinsic and extrinsic text quality. Qualitatively, we observe two stages of degradation. First, the emergence of attribute-congruent hallucinations, where the model fabricates facts to satisfy the steering objective (see Appendix D.12). This effect is non-uniform across attributes: content-heavy properties like sentiment are significantly more prone to hallucinations than stylistic properties like readability.

At even larger steering strengths, we observe a second stage of degradation manifesting as the generation of repetitive sequences of words strongly associated with the steering direction. This collapse is particularly pronounced for the toxicity steering vector, which exhibits the steepest decline in quality metrics (Figures 6 and 7). Unlike other attributes that degrade gradually, overriding safety training to induce toxicity frequently causes the model to devolve into incoherent loops of toxic keywords rather than fluent summaries, as illustrated in Table 12.

Finally, we find that larger models are consistently more robust to degradation, maintaining higher coherence and faithfulness at steering strengths that cause strong quality degradation in smaller models.

5.4 Comparing Steering to Prompting

We compare steering vectors with prompt engineering to evaluate their relative control strength and impact on text quality. For each target property, we designed encouraging, neutral, and discouraging prompt variations by appending specific behavioral directives, such as using simple language for readability, or emphasizing optimistic viewpoints for sentiment, to the standard base prompt (see Appendix B). Averaged results for Llama 1b are presented in Table 1; more detailed results can be found in Appendix D.2 and Appendix D.3. We find that steering offers stronger control than prompting for smaller models, but prompting and steering are on par for larger models. While both methods are more effective for larger models, prompting benefits even more from model scale. Furthermore, we find that prompting has only a negligible impact on text quality, as detailed in Appendix D.5.

Table 1: A comparison of mean metric values for the Llama 1b model on NEWTS shows that steering generally offers stronger control over summary properties than prompt engineering for small models. For topic and sentiment, a steering strength of $\lambda = 1$ matches or exceeds the effects of prompting, while $\lambda = 2$ produces an even larger effect. Prompting, however, is more effective at increasing readability complexity and has simplification effects similar to steering. The impact on toxicity is negligible for both methods, with the exception of strong steering ($\lambda = 2$), which also degrades text quality. Individual metric values are provided in Appendix D.2.

	Steering with strength λ		Prompting model for behavior			Steering with strength λ	
Behavior	$\lambda = -2$	$\lambda = -1$	Discourage	Neutral	Encourage	$\lambda = 1$	$\lambda = 2$
Topic	0.02 ± 0.0	0.10 ± 0.0	0.13 ± 0.0	0.14 ± 0.0	0.16 ± 0.0	0.16 ± 0.0	0.25 ± 0.0
Sentiment	-0.55 ± 0.3	-0.30 ± 0.4	-0.30 ± 0.3	-0.08 ± 0.5	0.27 ± 0.4	0.20 ± 0.5	0.79 ± 0.1
Readability	6.69 ± 3.5	6.52 ± 2.3	7.19 ± 3.6	6.00 ± 2.7	5.00 ± 2.1	4.94 ± 2.8	5.40 ± 5.7
Toxic	0.00 ± 0.0	0.00 ± 0.0	0.00 ± 0.0	0.00 ± 0.0	0.01 ± 0.0	0.00 ± 0.0	0.10 ± 0.0

5.5 Combining Steering and Prompting

A combined strategy of steering with prompting, where prompts are encouraging for $\lambda > 0$, neutral for $\lambda = 0$ and discouraging for $\lambda < 0$, leads to greater effect sizes.

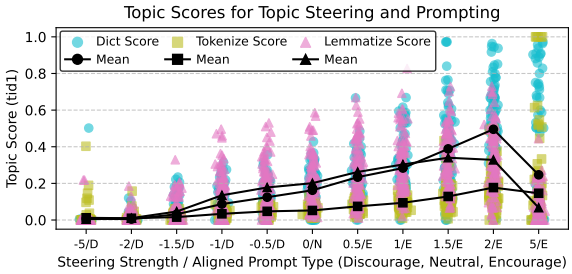


Figure 8: For Gemma 12b, combining steering and prompting provides stronger control over topical focus than either technique alone. Topical focus increases with larger λ values until text degradation reduces scores.

This combined approach yields substantially stronger control across all evaluated target behaviors and models. For instance, it more strongly influences topical focus (Figure 8) and achieves significant sentiment shifts with smaller steering values (Figure 11). Notably, the approach is also powerful enough to elicit toxic generations without causing a complete breakdown in text generation quality (Figure 9). Qualitative analysis further confirms this effectiveness; as shown in Table 2, combining steering with targeted prompting allows for precise modulation of readability, enabling the model to transition from dense, academic lexicon to simplified language while retaining the core narrative. Appendix D.7 provides a side-by-side comparison with steering-only results, illustrating this significant improvement in control.

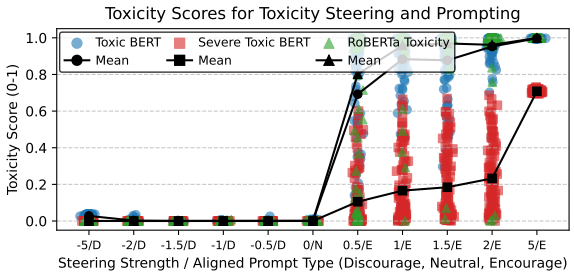


Figure 9: For Llama 3b, strong increases in toxicity at moderate λ values occur exclusively when combining prompting and steering.

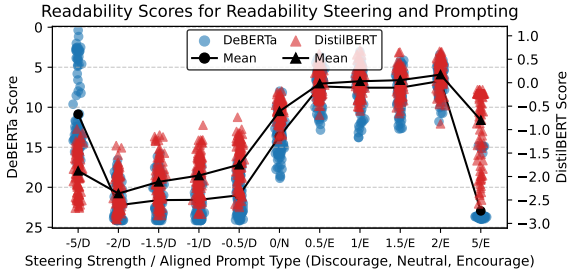


Figure 10: Combined steering and prompting impacts text readability more strongly than either method alone. In Llama 3b and for $\lambda > 2$, substantial text degradation causes different readability metrics to offer divergent assessments of complexity.

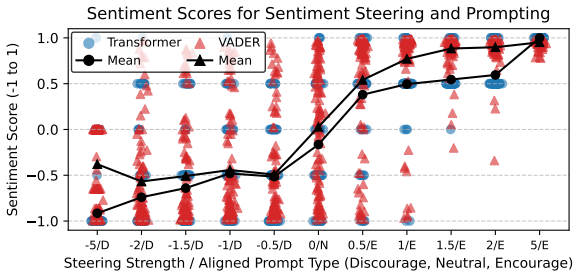


Figure 11: Combined steering and prompting achieves significant average sentiment changes from baseline (to approx. ± 0.5) with $\lambda = \pm 0.5$ for Gemma 12b. Steering alone requires $\lambda \approx \pm 1.5$ to achieve similar respective positive or negative shifts. This synergistic advantage diminishes for larger λ magnitudes.

Table 2: Example summaries for combined readability steering and prompting applied to the Llama-3.2-3B model, demonstrating varied outputs for the same source text. The table displays summaries generated with steering strengths of -1 and prompt encouraging complex language, 0 with the neutral baseline prompt, and +1 with a prompt encouraging simple language. Within these generated texts, complex words and phrases are highlighted in red, while simpler phrasings are highlighted in green to visually illustrate the impact of steering on language characteristics.

$\lambda = -1$ **+ complexity prompting**	$\lambda = 0$ **+ neutral prompting**	$\lambda = +1$ **+ simplicity prompting**
The United States is precipitating a global economic calamity, owing to its failure to devise a plan to augment its debt ceiling and avert default, thereby precipitating a calamitous event that will have far-reaching consequences for the global economy. As the World Bank President, Jim Yong Kim, has warned, the US's inaction will precipitate a cataclysmic collapse of confidence, leading to a precipitous decline in economic growth, and ultimately, a devastating impact on ...	The World Bank President, Jim Yong Kim, warned that the US is just five days away from causing a global economic disaster unless a plan is put in place to raise the nation's debt limit and avoid default. The US debt ceiling deadline is approaching and Treasury Secretary Jacob Lew has warned that the government will exhaust its borrowing authority on Thursday. The World Bank ...	The World Bank says the US is close to causing a big economic problem if the politicians don't fix the debt limit. The US needs to borrow money to pay its bills, but the government is running out of money and might not be able to pay its debts. The World Bank says this could make interest rates go up, and that could make it harder for people to borrow money and for the economy to grow ...

5.6 Summary Quality Degradation for Combined Steering and Prompting

Combining prompting and steering not only amplifies the effect size, but also undesirable quality degradation of the generated summaries. Details can be found in the Appendix D.9. In general, the combination of both techniques provides the most favorable trade-off between efficacy and quality.

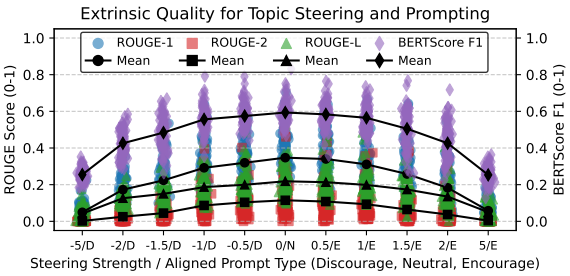


Figure 12: For Qwen 3b, combining steering and prompting for topical focus impacts extrinsic quality more negatively than steering alone, especially for steering magnitudes $|\lambda| > 1$.

5.7 Conditional Steering

We apply steering vectors conditionally to a model's activations. The motivation is that if the model is already generating text that aligns with a target behavior, further steering is unnecessary and might degrade text quality. To implement this, at each generation step, the current activation is projected onto the difference of means line, the vector connecting the mean activations of negative and positive training samples. If this projected value is greater than a threshold of 1, it signifies that the model's current state is already more aligned with the target behavior than the average positive training example. In this case, no steering is applied for that specific token. This projection is done at every generation step, so in a sequence, some token positions are steered and others are not. We find this method meaningfully reduces unintended side effects on text quality, but it also weakens the overall control of the target behavior (see Table 3). Conditional steering likely prevents activations from going out-of-distribution in cases where cumulative steering at prior generation steps has already sufficiently biased the current context.

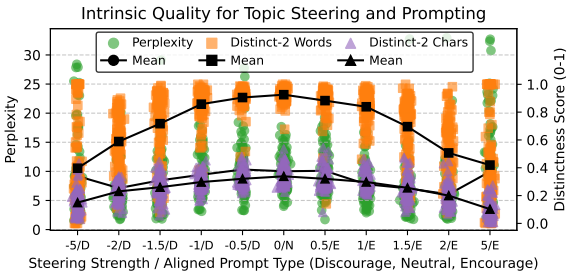


Figure 13: Similarly, intrinsic quality for Qwen 3b degrades for $|\lambda| > 1$. At lower λ values, however, this strategy enables stronger topical focus than steering alone while causing only minimal degradation.

Table 3: A comparison of mean metric values for the Gemma 4b model on the SAMSum, NEWTS, and arXiv datasets. The table shows the intended effect for steering the target behaviors alongside the unintended text quality degradation, for a conditional steering threshold of $\tau = 1$. Values are computed relative to unconditional steering with $\lambda = 1$. Conditional steering meaningfully reduces text degradation, but also weakens control.

	SAMSum		NEWTS		arXiv	
Behavior	Effect	Degradation	Effect	Degradation	Effect	Degradation
Topic	N/A	N/A	0.01 ± 0.0	-0.1 ± 0.1	N/A	N/A
Sentiment	0.0 ± 0.1	-0.1 ± 0.0	-0.12 ± 0.3	-0.05 ± 0.0	-0.09 ± 0.2	-0.1 ± 0.0
Readability	-0.25 ± 0.2	-0.1 ± 0.2	-0.18 ± 0.3	-0.07 ± 0.1	-0.41 ± 0.3	-0.08 ± 0.2
Toxic	-0.01 ± 0.0	-0.01 ± 0.0	-0.01 ± 0.0	-0.01 ± 0.0	-0.00 ± 0.0	-0.01 ± 0.0

6 Discussion

Q1. How effective are steering vectors for controlling free-form summarization? Our findings demonstrate the effectiveness of difference-of-means steering vectors for controlling relevant text properties during free-form abstractive summarization. Specifically, we find that steering effectively controls topical focus, sentiment, and readability but this control inherently involves an efficacy-quality trade-off: higher steering strengths achieve greater control at the cost of significant degradation in both intrinsic and extrinsic summary quality.

Steering against the model's safety training for toxicity proved challenging for all models. Suppressing toxicity via negative steering yielded no quantitative changes due to a floor effect, as the baseline toxicity of our datasets (SAMSum, NEWTS, arXiv) is negligible. Conversely, when inducing toxicity via positive steering to probe the robustness of safety alignment, we found that steering alone failed to reliably produce coherent toxic content. The high positive steering strengths needed to override the model's safety alignments severely degraded text quality, often reducing the output to repetitive sequences of toxic words. This confirms that while the steering vector successfully captures the direction of toxicity, the model's safety training successfully resists generating coherent summaries with the property that was suppressed during model training.

Q2. What are the side effects of steering on summary quality? A central finding of this study is the existence of an efficacy-quality trade-off. While higher steering strengths achieve greater control over the target attribute, they do so at the cost of a significant degradation in both intrinsic (coherence, fluency and repetitions) and extrinsic (summary faithfulness and hallucinations) summary quality. We observed that strong steering causes the model to prioritize the steering objective over source fidelity, leading to attribute-congruent hallucinations. For example, steering for positive sentiment occasionally induced the fabrication of positive events not present in the source text, while negative steering invented distinct negative details to fit the targeted tone. This trade-off underscores that practitioners cannot increase control indefinitely without consequence. The degradation in quality at high steering strengths appears to be a result of pushing the model's internal activations into out-of-distribution states, which degrades the model's ability to generate coherent and faithful summaries. We also observed that toxicity steering influenced summary sentiment.

Q3. How does steering compare to prompting for text control? When compared to conventional prompt engineering, steering vectors offer a different set of trade-offs. Our results show that steering generally offers stronger control over text properties compared to prompting, particularly for smaller models where prompting is less effective. Prompting, however, preserves summary quality much better than steering, making it preferable when maintaining high coherence is paramount and only moderate control is needed. Furthermore, prompting remains the most practical method for arbitrary, ad-hoc tasks where creating a specialized dataset to train a steering vector is impractical. Consequently, steering excels at controlling broad, pervasive text properties like style or sentiment, but it does not replace prompting for complex, specific instruction following.

Q4. How effective is a hybrid steering and prompting approach? Combining steering vectors with instructive prompts emerged as the most promising strategy in our evaluation. This hybrid approach yielded the strongest control over text

attributes, often achieving a greater effect with moderate steering strengths than steering could alone, even at high strengths. Consequently, the hybrid method achieved the most favorable efficacy-quality trade-off. By leveraging the complementary strengths of both techniques—the precise control of steering and the quality-preserving nature of prompting—it is possible to exert strong control while mitigating some of the quality degradation. However, it is important to note that even with this combined approach, the use of very large steering strengths still leads to a substantial decline in text quality.

Q5. Can conditional steering mitigate steering side effects while maintaining control efficacy? We find that conditional steering successfully mitigates steering side effects but with reduced steering effect size. Adjusting the steering strength λ dynamically, based on the model's current activations is a promising approach to find a better control-quality tradeoff.

Q6. How do model scale and architecture impact steering? Increased model scale improves steering efficacy by strengthening control while simultaneously mitigating text quality degradation. This is likely because the additional layers in larger models provide a greater capacity to absorb the steering intervention without disrupting text generation. Furthermore, results were broadly consistent across Llama, Qwen, and Gemma architectures, indicating that steering is a robust and generalizable technique for transformer-based models.

6.1 Future Work

The observed trade-off between control efficacy and text quality motivates a systematic study comparing different steering methods on free-form generation tasks. This would entail comparing difference-of-means vectors (Rimsky et al., 2024b) with (Hendel et al., 2023, Task Vectors), (Todd et al., 2024, Function Vectors) and (Singh et al., 2024, MiMiC) to determine which offers the best balance between control and quality preservation.

Furthermore, it would be beneficial to compare steering methods against a broader range of established control methods from trainable decoding mechanisms, to fine-tuning. A direct comparison would clarify the specific advantages and disadvantages of each approach, helping practitioners determine whether steering offers a more efficient or effective solution than retraining for their specific use case.

Another important area for future exploration is the application of steering vectors in multiple-attribute controllable summarization. This would involve developing and applying methods to steer multiple text properties simultaneously. This approach could present new challenges related to vector composition, possible interference between steering directions, and managing cumulative impacts on text quality.

6.2 Conclusion

Steering vectors represent an effective but lightweight method for adapting large-scale foundation models to user preferences at inference time. We find that difference of means steering vectors are effective at controlling text properties in free-form adaptive summarization, but their use is governed by a critical trade-off between control efficacy and text quality. We observe that large steering strengths consistently induce degenerate repetition and factual hallucinations in generated summaries. The combination of steering and prompting provides the most effective balance between control and quality. Our work points toward hybrid methods as a promising path for efficiently and robustly aligning LLM behavior with user preferences in complex, real-world applications.

Limitations

Our findings should be interpreted within the context of several methodological constraints. First, while we evaluate thirteen models across three families, our study is restricted to dense transformer architectures between 0.6B and 70B parameters. Second, our evaluation relies on automated metrics. While we validate these against LLM-based judges, they serve as proxies and cannot fully capture the nuances of human preference, particularly regarding subtle factual hallucinations induced by high steering strengths. Third, we focus exclusively on difference-of-means steering vectors applied to heuristically selected layers. We did not perform an exhaustive hyperparameter search for optimal injection layers, nor did we compare against other activation engineering techniques or fine-tuning based alternatives. Finally, our experiments are limited to English-language datasets; the efficacy of steering vectors in multilingual or low-resource settings remains an open question for future work.

Author Contributions

J.B. conceptualized the study, implemented and performed all experiments and analysis, and wrote the initial draft of the manuscript. C.E. provided senior supervision, resources, and feedback on the final version of the paper. S.A.B. provided regular supervision throughout the project, contributed to the experimental design, assisted with the rebuttal process, and refined the manuscript.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback which helped to improve the manuscript. This research utilized compute resources at the Tübingen Machine Learning Cloud, DFG FKZ INST 37/1057-1 FUGG.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Preprint*, arXiv:2406.11717.

Seyed Ali Bahrainian, Jonathan Dou, and Carsten Eickhoff. 2024. Text simplification via adaptive teaching. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6574–6584, Bangkok, Thailand. Association for Computational Linguistics.

Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. 2022. NEWTS: A corpus for news topic-focused summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 493–503, Dublin, Ireland. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Sofia Blinova, Xinyu Zhou, Martin Jaggi, Carsten Eickhoff, and Seyed Ali Bahrainian. 2023. SIMSUM: Document-level text simplification via simultaneous summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9927–9944, Toronto, Canada. Association for Computational Linguistics.

Joschka Braun, Carsten Eickhoff, and Seyed Ali Bahrainian. 2025a. Beyond multiple choice: Evaluating steering vectors for adaptive free-form summarization. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. 2025b. Understanding (un)reliability of steering vectors in language models. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Joschka Braun, Dmitrii Krasheninnikov, Usman Anwar, Robert Kirk, Daniel Chee Hian Tan, and David Scott Krueger. 2024. A sober look at steering vectors for llms. AI Alignment Forum. Publication Date: 2024-11-23.

Joschka Braun, Bálint Mucsányi, and Seyed Ali Bahrainian. 2025c. Logit reweighting for topic-focused summarization. *Preprint*, arXiv:2507.05235.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Madeline Brumley, Joe Kwon, David Krueger, Dmitrii Krasheninnikov, and Usman Anwar. 2024. Comparing bottom-up and top-down steering approaches on in-context learning tasks. *Preprint*, arXiv:2411.07213.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. 2020. Residual energy-based models for text generation. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333, Singapore. Association for Computational Linguistics.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Kai Konen, Sophie Freya Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style Vectors for Steering Generative Large Language Models. In *European Chapter of the ACL: (EACL) 2024*, St Julians, Malta.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, 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. In *Advances in Neural Information Processing Systems*.

Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, and David Krueger. 2024. Towards reliable evaluation of behavior steering interventions in llms. In *MINT: Foundation Model Interventions*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024a. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024b. Steering Llama 2 via Contrastive Activation Addition. *arXiv preprint*. ArXiv:2312.06681 [cs].

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. 2024. Mimic: Minimally modified counterfactuals in the representation space. *CoRR*, abs/2402.09631.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting Latent Steering Vectors from Pretrained Language Models. In *Findings of the Association for Computational Linguistics: ACL*

2022, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.

Daniel Chee Hian Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. Analysing the generalisation and reliability of steering vectors. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Curt Tigges, Oskar J. Hollinsworth, Atticus Geiger, and Neel Nanda. 2024. Language models linearly represent sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 58–87, Miami, Florida, US. Association for Computational Linguistics.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2024. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*.

NLP Town. 2023. bert-base-multilingual-uncased-sentiment (revision edd66ab).

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation Addition: Steering Language Models Without Optimization. *arXiv preprint*. ArXiv:2308.10248 [cs] version: 3.

Ashok Urlana, Pruthwik Mishra, Tathagato Roy, and Rahul Mishra. 2024. Controllable text summarization: Unraveling challenges, approaches, and prospects - a survey. In *ACL (Findings)*, pages 1603–1623.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. 2025. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *Preprint*, arXiv:2403.02901.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

A Datasets

To rigorously evaluate steering vectors and the the trade-off between control efficacy and text quality, we selected datasets that represent a diverse spectrum of linguistic complexity, input length, and domain formality.

A.1 SAMSum Dataset

We utilize the SAMSum dataset (Gliwa et al., 2019) to represent the domain of short, informal, and conversational text. The dataset consists of approximately 16,000 messenger-like conversations accompanied by human-written abstractive summaries. These dialogues were constructed by linguists to emulate real-life instant messaging, featuring natural conversational phenomena such as slang, emoticons, typos, and overlapping threads. SAMSum serves as a critical stress test for steering vectors due to its high level of linguistic "noise". The conversations are short, distributed uniformly across lengths of 3 to 30 utterances.

A.2 NEWTS Dataset

The NEWTS (NEWs Topic-based Summarization) dataset (Bahrainian et al., 2022) is based on the CNN/DailyMail dataset (Nallapati et al., 2016) and consists of 2400 training and 600 test samples. Each sample provides a source article and two human-written reference summaries, each focussed on either one of the two most prominent topics in the article. The dataset is specifically curated for topic-focused summarization and has 50 unique topics. The majority of articles are between 250 and 1000 tokens long and summaries are between 25 and 200 tokens (see Table 4). For our evaluation, NEWTS represents the standard for polished, medium-length, journalistic English. It serves as a middle ground between the casual nature of SAMSum and the technical density of arXiv. Because NEWTS was designed to benchmark topic-focused generation, it provides an ideal testbed for evaluating our topical steering vectors.

A.3 arXiv Dataset

To evaluate our method on the most challenging end of the linguistic spectrum, we utilize the arXiv dataset introduced by Cohan et al. (2018). This dataset consists of scientific papers collected from the arXiv preprint server, where the task is to generate the paper's abstract based on the full body of the text. With an average document length of approximately 6,000 tokens, the inputs are nearly an order of magnitude longer than the average NEWTS article. This provides a rigorous test for the robustness of steering vectors, allowing us to analyze whether the steering influence persists throughout long-context generation.

A.4 Steering Vector Training Datasets

Our steering vector training datasets were constructed by prompting a large language model to synthesize contrastive samples. For each behavioral property, we used 100 contrastive samples for training, as preliminary experiments showed convergence in steering vector direction beyond this threshold. The full datasets (500 samples each for Sentiment, Toxicity, and Readability) are publicly released under the Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) license, while the accompanying codebase is provided under the MIT License.

A.4.1 Generation and Quality Control

The Sentiment, Toxicity, and Readability datasets were generated using Claude 3 Sonnet (version 20240229) with 1-shot prompting to produce contrastive pairs differing primarily in the target attribute. To ensure topical diversity, generation was distributed across 20 categories: Technology and Innovation, Food and Cuisine, Nature and Environment, Sports and Athletics, Health and Medicine, Arts and Culture, Music and Sound, Media and Entertainment, Weather and Climate, Education and Learning, Politics and Government, Emotions and Feelings, Fashion and Clothing, Travel and Transportation, Law and Justice, Science and Research, Architecture and Design, Family and Relationships, Business and Finance, and Philosophy and Ethics.

Quality control followed a consistent pipeline: (i) initial generation of 1,000 candidate samples, (ii) manual review and deduplication via token-overlap matching, (iii) removal of outliers by length, and (iv) random subsampling to obtain the final 500 samples.

A.4.2 Dataset Specifications

Sentiment The dataset contains 500 single-sentence pairs, where each pair consists of one sentence expressing positive sentiment and one expressing negative sentiment on the same topic. Sentences were constrained to be grammatically complete, avoid references to specific products or media, and maintain parallel structure across polarities.

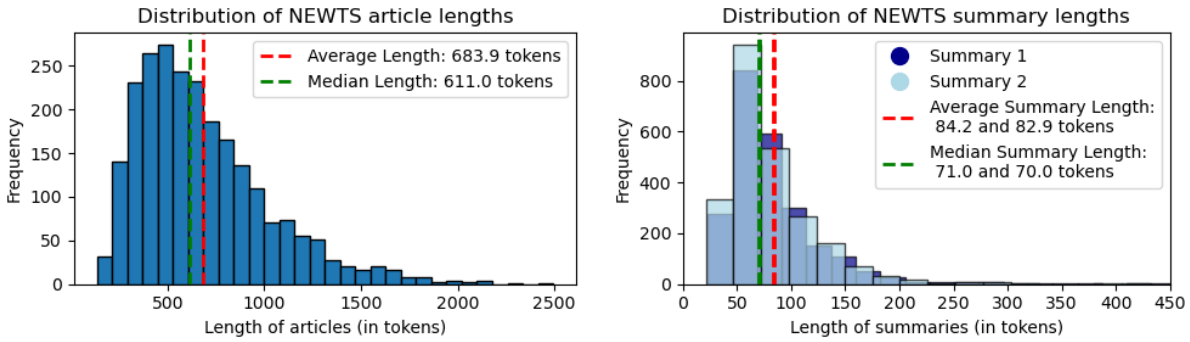


Figure 14: NEWTS article length and summary length distributions for Llama 3 tokenizer.

Table 4: An example from the NEWTS dataset. The source article discusses a U.S. debt ceiling standoff and its global economic implications. Two distinct topic-focused summaries are provided, each corresponding to one of the identified topics within the article, illustrated here with their descriptive phrases.

Article Snippet: The president of the World Bank on Saturday warned the United States was just 'days away' from causing a global economic disaster unless politicians come up with a plan to raise the nation's debt limit and avoid default. 'We're now five days away from a very dangerous moment. I urge US policymakers to quickly come to a resolution before they reach the debt ceiling deadline ...
Topic 1 (tid1): 175 **Topic Description:** This topic is about the senate and congress, congressional pressure, calling one's representative's office, informing a Senate committee, lawmakers setting the record straight, the staffer to the Democratic senator, and federal employee benefits. **Summary 1 (Focused on Topic 1):** The leader of the World Bank urged the US to take action before the borrowing deadline. The US Congress needed to come to an agreement to raise the borrowing limit, as the UD treasury secretary had stated his authority had reached its limits in the matter. Republicans shot down the Democratic proposal to increase the borrowing limit, putting a federal default at risk that would affect the global economy.
Topic 2 (tid2): 110 **Topic Description:** This topic is about economic growth involving billion dollar figures showing that the economy is growing as expected globally. **Summary 2 (Focused on Topic 2):** The US economy will be a driving factor in the world economy for many coming years, the stability and growth of the US economy is crucial on a global scale. The US had reached its debt ceiling and many world banks and leaders grew concerned. Having failed to reach an agreement, the US will be unable to virtue any further, risking federal default and collapse of the worlds economies.

Toxicity The dataset contains 500 multi-sentence pairs contrasting toxic and non-toxic language. Toxic samples exhibit harmful language, hostile tone, or offensive content, while non-toxic counterparts convey equivalent information in a neutral or constructive manner.

Readability The dataset contains 500 pairs of text summaries at contrasting complexity levels. Generation followed a two-step process: the model first produced a short fictional article on a given topic, then generated both a simple summary (accessible vocabulary, short sentences) and a sophisticated summary (advanced vocabulary, complex sentence structure) of the same content. This was repeated 50 times per topic category to ensure coverage.

Topical Focus Unlike the above datasets, the Topical Focus steering vectors utilize topic descriptions from the NEWTS dataset (Bahrainian et al., 2022). For each of the 50 topics, we constructed a steering vector using 49 contrastive pairs, where each pair contrasts the target topic description against one of the remaining 49 topic descriptions.

B Prompt Variations

B.1 Prompt Design for Article Summarization

The system for generating article summarization prompts employs a structured approach, ensuring flexibility and control over the summarization output. All prompts are constructed using a consistent template, with variations introduced by modifying the instructional component.

B.1.1 Core Prompt Structure

The foundational structure for every prompt is defined by the following template:

Prompt template

```
{instruction}
Article:
{article}
Summary:
```

This template consists of three components:

1. **[Instruction Block]**: Represented by `{instruction}`, this section contains the specific directives given to the language model. Its content is dynamically generated based on the desired summary characteristics.

2. **[Article Placeholder]**: Denoted by `{article}`, this is where the actual text of the article to be summarized is inserted.

3. **[Summary Elicitation Cue]**: The literal string `"\nSummary:\n"` serves as a cue, guiding the model to generate the summary following this marker.

Variations in the summarization task are achieved by altering the content of the **[Instruction Block]**. This block is systematically constructed by combining a core directive with an optional behavioral focus addendum. The **[Instruction Block]** begins with a **[Core Directive]**, which is constant across all prompt types:

```
"Write a three sentence summary
of the article"
```

To tailor the summary, a **[Behavioral Focus Addendum]** can be appended to this **[Core Directive]**. This addendum specifies the particular aspect (e.g., topic, sentiment, readability) the summary should emphasize. Finally, a period is appended to the combined instruction before it is placed into the `{instruction}` slot of the template. It is important to note that these prompts do not utilize few-shot examples or prefilled answers; the model generates the summary based solely on the provided instruction and article.

B.1.2 Prompt Variations

The system implements five main categories of prompts, achieved by varying the **[Behavioral Focus Addendum]** within the **[Instruction Block]**:

1. **Neutral Summary Prompt**:

- **Formation**: The **[Instruction Block]** consists solely of the **[Core Directive]**. No **[Behavioral Focus Addendum]** is included.
- **Instruction Text**: `"Write a three sentence summary of the article."`
- **Purpose**: To generate a general, unbiased three-sentence summary of the article.

2. **Topic-Focused Summary Prompt**:

- **Formation**: A **[Behavioral Focus Addendum]** is appended to the **[Core Directive]** to steer the summary towards a specific subject.
- **Example Addendum**: `" focusing on the topic related to: {topic_description}"`, where `{topic_description}` is a comma-separated list of keywords defining the target topic (e.g., `"climate change, renewable energy, policy"`).
- **Instruction Text Example**: `"Write a three sentence summary of the article focusing on the topic related to: climate change, renewable energy, policy."`
- **Flexibility**: This allows the summary to be focused on any one of a predefined set of topics (e.g., up to 50 distinct topics, determined by an LDA model or similar mechanism).

3. **Sentiment-Focused Summary Prompt**:

- **Formation**: The **[Behavioral Focus Addendum]** guides the summary to adopt a specific emotional tone. This is a binary option.
- **Variations**:
  - *Positive Sentiment*: The addendum encourages highlighting favorable outcome and optimistic viewpoints. Example addendum: `" emphasizing positive outcomes and optimistic viewpoints"`.

  - *Negative Sentiment*: The addendum encourages emphasizing negative consequences and critical perspectives. Example addendum: `" emphasizing negative consequences, criticisms and concerns"`.
- **Instruction Text Example (Positive)**: `"Write a three sentence summary of the article emphasizing the positive outcomes, optimistic viewpoints, or favorable details presented in the article."`

4. **Toxicity-Focused Summary Prompt**:

- **Formation**: The **[Behavioral Focus Addendum]** controls the presence or absence of toxic language in the summary. This is a binary option.
- **Variations**:
  - *Encouraging Toxicity*: The addendum instructs the model to use toxic language. Example addendum: `" using toxic and harmful language"`.
  - *Avoiding Toxicity*: The addendum instructs the model to refrain from toxic language. Example addendum: `" while avoiding any toxic or harmful language"`.
- **Instruction Text Example (Avoiding Toxicity)**: `"Write a three sentence summary of the article while avoiding any toxic or harmful language."`

5. **Readability-Focused Summary Prompt**:

- **Formation**: The **[Behavioral Focus Addendum]** adjusts the linguistic complexity of the summary. This is a binary option.
- **Variations**:
  - *Encouraging Simplicity*: The addendum promotes the use of simple, easily understandable language. Example addendum: `" using simple and easy to understand language"`.
  - *Encouraging Complexity*: The addendum promotes the use of sophisticated and complex language. Example addendum: `" using complex and sophisticated language"`.
- **Instruction Text Example (Encouraging Simplicity)**: `"Write a three sentence summary of the article using simple and easy to understand language."`

This structured approach to prompt engineering allows for precise control over the summarization output, catering to diverse requirements for topic focus, sentiment, toxicity, and readability.

C Evaluation of Summaries

We evaluate generated summaries across six key dimensions: *intrinsic quality* based on text characteristics, *extrinsic quality* against reference summaries, *topical focus* relative to predefined topics, *sentiment* polarity, *toxicity* and *readability*. For robustness, we measure two to four metrics for each text property.

C.1 Intrinsic Quality Evaluation

Intrinsic quality, assessing the linguistic quality and fluency of the generated text without relying on reference summaries, is evaluated to measure undesirable generation artifacts.

Perplexity (PPL): Perplexity measures how well a pre-trained language model can predict the generated text sequence. A lower perplexity score generally indicates higher fluency and text that is more statistically likely according to the language model (Bengio et al., 2000).

Bigram Repetition (Distinct-2 Word): Distinct-2 Word measures textual diversity and penalizes unnatural word repetition. It is calculated as the ratio of unique word bigrams to the total number of bigrams in the generated text. Lower Distinct-2 scores indicate higher repetition, which often correlates negatively with human-annotated quality (Li et al., 2016).

Character Bigram Repetition (Distinct-2 Char): Distinct-2 Char assesses fine-grained textual diversity and penalizes character sequence repetition. This metric is calculated as the ratio of unique character bigrams to the total number of character bigrams. It is particularly useful for texts without clear word separation and for identifying various forms of text degradation; lower scores signify increased character bigram repetition and potential quality issues.

C.2 Extrinsic Quality Evaluation

To evaluate extrinsic quality, we measure the similarity and faithfulness of generated summaries to their respective NEWTS reference summaries using the following metrics:

ROUGE Score: Recall-Oriented Understudy for Gisting Evaluation (ROUGE) includes three variants that quantify the overlap between a candidate summary c and a reference r. ROUGE-1 and ROUGE-2 respectively assess unigram and bigram overlap considering recall, precision and F_1, while ROUGE-L measures the longest common subsequence. Collectively, these metrics capture content fidelity, fluency and sequence-level coherence (Lin, 2004).

BERTScore: BERTScore (Zhang* et al., 2020) leverages contextual embeddings from the pre-trained transformer model to compute semantic similarity between two text distributions. This makes the metric robust against paraphrasing, a key advantage over ROUGE scores. For our evaluation, we employ the **'BERTScorer'** class with the `microsoft/deberta-xlarge-mnli` model (He et al., 2021), selected for its strong correlation with human evaluations of semantic content.

C.3 Topical Focus Evaluation

To evaluate the alignment of generated summaries with the intended topics, we utilize three methods to quantify topical focus:

Lemmatization-Based Scoring: This method processes the generated text by lemmatizing words to their canonical forms. Using the LDA model, it matches these lemmas against the lemmas of the top topic words identified for the relevant topic. The topical focus score is then calculated as the weighted presence of these lemmas in the summary, normalized by the total weight of all top topic lemmas.

Tokenization-Based Scoring: This approach tokenizes the summary using the *bert-base-multilingual-uncased* tokenizer. The score represents the proportion of tokens in the summary that match the token IDs derived from the top words of the target LDA topic, providing a direct measure of topical vocabulary usage at the sub-word level.

Dictionary-Based Evaluation: This method employs a bag-of-words representation for the summary, utilizing the Gensim dictionary associated with the LDA model. The LDA model infers a topic distribution for the summary, and the score reflects the computed prevalence of the target topic within this distribution.

C.4 Topic Representations

The 50 latent topics derived from the LDA model in the NEWTS dataset (Bahrainian et al., 2022) provide a compelling target for steering language models. Unlike binary qualities such as sentiment or toxicity, these topics represent more nuanced,

Table 5: Table illustrating different types of topic representations with examples.

Representation Type	Representation
words	"children", "child", "parents", "birth", "born", "kids", "families", "mother", "family", "care", "daughter", "young", "girl", "syndrome", "adults",
n-grams	"children and parents", "families with children", "having kids", "giving birth", "she became a mother", "baby was born"
descriptions	"This topic is about having kids, becoming a mother, giving birth, children and their parents, and families with children when a baby is born."
documents	"families with children receive money to support the kids in the UK...", "Children with special needs were mentioned in a political campaign...", "Only half of British children live with both parents..."

multi-faceted concepts that can be understood through various representations, making them an interesting challenge. Steering topical focus is also practically relevant, for instance, when summarizing information for a particular stakeholder or expert, as it allows for the selection of content most important to that specific reader. Topic representations are presented in Table 5.

C.5 Sentiment Evaluation

To evaluate the sentiment expressed in the generated summaries, we use two approaches:

Lexicon-Based Analysis (VADER): We incorporate VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto and Gilbert, 2014), a lexicon and rule-based sentiment analysis tool. VADER provides multiple scores, including a normalized compound score ranging from -1 (most negative) to +1 (most positive), effective at capturing sentiment intensity and negation.

Transformer-Based Analysis: We leverage a pre-trained transformer model fine-tuned for sentiment classification: *nlptown/bert-base-multilingual-uncased-sentiment* (Town, 2023). We renormalize the model output to -1 to 1.

C.5.1 Toxicity Evaluation

Abstractive summaries must not reproduce hateful, harassing, or threatening language. We therefore measure toxicity for every generated summary with two Transformer classifiers. Toxicity is also a challenging property for steering experiments, as language models typically undergo extensive post-training alignment to curb the generation of such content, making any residual or induced toxicity a notable outcome to control.

Toxic-BERT Toxic-BERT is a BERT-base model fine-tuned to predict the probabilities for eight labels (*toxic*, *severe_toxic*, *obscene*, *threat*, *insult*, *identity_attack*, *sexual_explicit*, *non_toxic*) (Devlin et al., 2019). We use the *toxic* and *severe_toxic* logits, normalised to the range $[0, 1]$, as separate indicators of surface-level and extreme toxicity.

RoBERTa toxicity classifier This classifier distils RoBERTa-base (Liu et al., 2019), producing a binary toxicity score between $[0, 1]$. Its more conservative calibration complements Toxic-BERT's multi-label view.

C.6 Readability Evaluation

Readability and language complexity are especially important text properties. Steering for readability is particularly relevant as it enables the generation of text summaries personalized to a user's specific comprehension level, for instance, matching their educational background or literacy skills. We therefore quantify the readability of each summary with two regression models.

DistilBERT fine-tuned for readability The DistilBERT variant (Sanh et al., 2020) was fine-tuned for readability and produces a continuous score in $[-5, 5]$ with higher values signifying high readability and negative values signifying low readability.

DeBERTa-V3 Fine-tuned version of DeBERTa-V3 (He et al., 2023) to predict U.S. grade levels (1–18). Therefore low scores correspond to simple text, and high scores to complex texts.

C.7 Validation of Automated Metrics

While human or LLM-based evaluation is often considered the gold standard for assessing perceptual qualities, we prioritized established automated metrics to ensure invariance to model prompting

Table 6: **Validation of Automated Metrics via LLM-as-a-Judge.** Scores are averaged over $N = 250$ samples using `gpt-5-mini`. The table compares human-aligned LLM scores (0–1) against the automated reference metrics used in this work across varying steering strengths (λ).

λ	**LLM Extrinsic** (Faithfulness)	BERTScore (Ref)	**LLM Intrinsic** (Fluency)	Distinct-2 (Ref)	**LLM Readability** (Simplicity)	DistilBERT (Ref)
-5.0	0.33	0.26	0.24	0.42	0.02	-1.8
-2.0	0.68	0.44	0.38	0.60	0.06	-2.4
-1.5	0.72	0.49	0.72	0.73	0.14	-2.1
-1.0	0.89	0.54	0.94	0.86	0.17	-1.9
-0.5	0.92	0.56	0.97	0.91	0.41	-1.7
0 (Base)	**0.93**	**0.60**	0.98	**0.93**	0.80	-0.6
0.5	0.91	0.57	**0.99**	0.88	0.94	-0.1
1.0	0.92	0.54	0.93	0.84	0.92	0.0
1.5	0.84	0.51	0.69	0.70	0.94	0.0
2.0	0.69	0.45	0.45	0.53	0.84	0.2
5.0	0.37	0.26	0.27	0.44	0.53	-0.8

and to facilitate large-scale sweeps over multiple datasets and steering strengths λ.

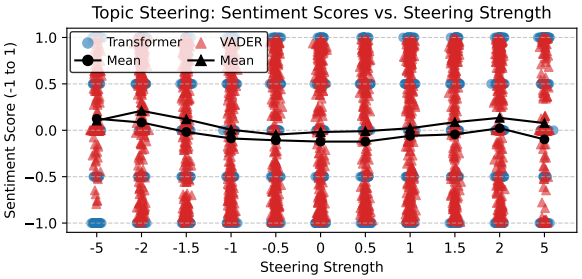
To validate that these automated metrics serve as reliable proxies for perceived quality, we conducted a correlation study using `gpt-5-mini-2025-08-07` as an LLM judge. We re-evaluated a subset of summaries ($N = 250$) across the steering spectrum, instructing the model to rate three dimensions on a continuous 0–1 scale:

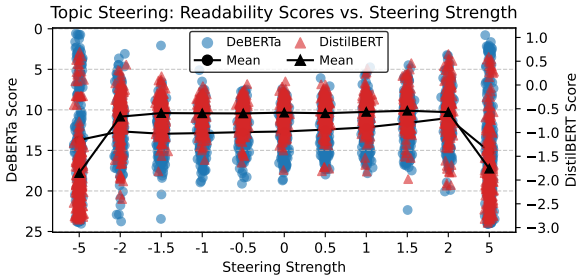
- **Extrinsic Quality:** Source-faithfulness and accuracy.
- **Intrinsic Quality:** Coherence, grammar, and fluency.
- **Readability:** Text simplicity and accessibility (where 1 = simple/clear, 0 = complex).

The results, presented in Table 6, demonstrate that our automated metrics strongly correlate with LLM-based judgments. Both evaluation methods reveal that mild steering ($|\lambda| < 1$) maintains quality, while high magnitudes ($|\lambda| > 2$) cause significant degradation. Furthermore, the correlation between DistilBERT and LLM-assessed readability confirms that our steering effectively modifies text complexity as intended.

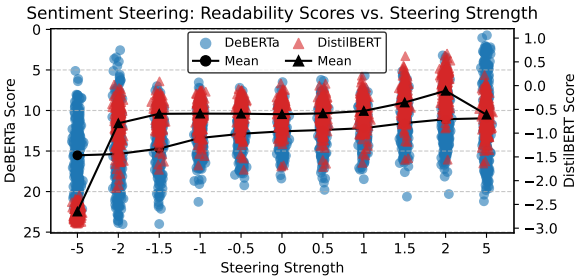
D Extended Results

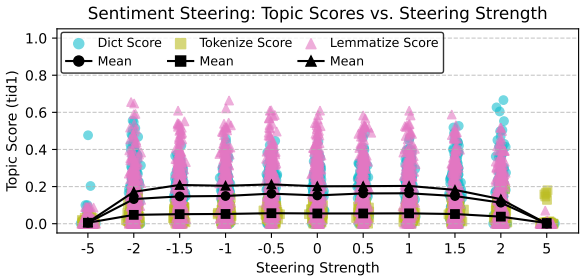
D.1 Steering Vectors do not change unrelated properties, except for toxicity impacting sentiment



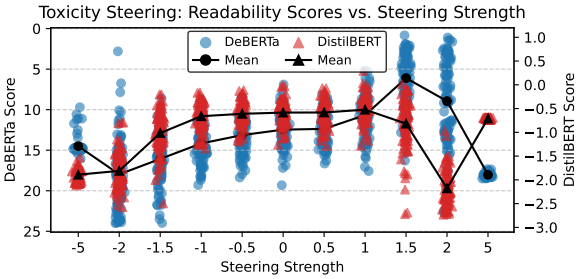


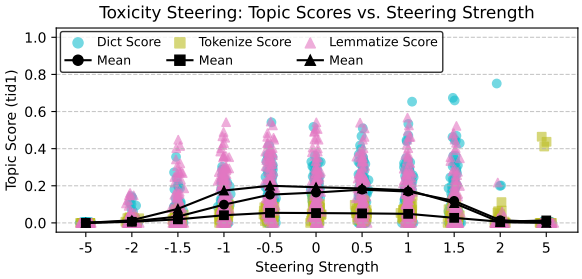
(a) In both cases, topic steering neither changes sentiment scores nor readability scores in a meaningful way. Readability scores only change once text degradation is significant for steering strengths larger than 2.



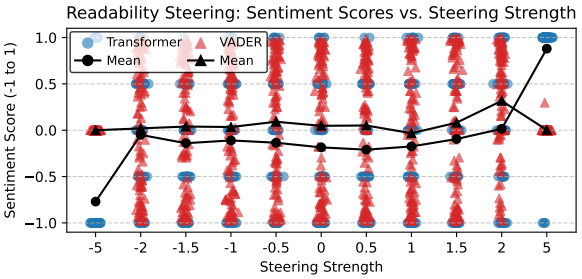


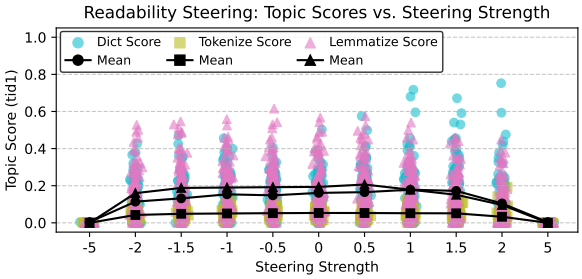
(b) Sentiment steering does not meaningfully impact readability or topic scores, except when generation quality degrades for $|\lambda| > 2$





(c) Steering for toxicity does not impact readability or topic scores for $\lambda \leq 1$. For $\lambda > 1$ strengths text quality degrades and scores vary.





(d) Except for very large steering strengths, readability steering does not impact unrelated text properties.

Figure 15: Steering for one text property does not impact other text properties, with the exception of toxicity steering impacting sentiment shown in Figure 4. Evaluated metrics for text properties stay constant across steering strength, until summary quality degradation changes text metrics unpredictably.

D.2 Comparing Steering and Prompt Engineering

Table 7: Mean metric values comparing control of summary properties via steering (λ) versus prompt engineering on NEWTS and Llama 1b. Steering generally offers stronger control than prompting. For topic and sentiment, $\lambda = 1$ matches or exceeds prompting effects, while $\lambda = 2$ has an even larger effect. Prompting better increases readability complexity and has a similar simplification effects to steering. Effects on toxicity are negligible for both methods, except for $\lambda = 2$ which also degrades text quality.

	Steering with strength λ		Prompting model for behavior			Steering with strength λ	
Behavior	$\lambda = -2$	$\lambda = -1$	Discourage	Neutral	Encourage	$\lambda = 1$	$\lambda = 2$
Topic							
dict	0.02 ± 0.0	0.11 ± 0.0	0.15 ± 0.0	0.16 ± 0.0	0.19 ± 0.0	0.21 ± 0.0	0.39 ± 0.0
stem	0.02 ± 0.0	0.10 ± 0.0	0.13 ± 0.0	0.13 ± 0.0	0.14 ± 0.0	0.14 ± 0.0	0.18 ± 0.0
lemmatize	0.04 ± 0.0	0.16 ± 0.0	0.21 ± 0.0	0.21 ± 0.0	0.23 ± 0.0	0.23 ± 0.0	0.29 ± 0.0
tokenize	0.01 ± 0.0	0.04 ± 0.0	0.06 ± 0.0	0.06 ± 0.0	0.07 ± 0.0	0.07 ± 0.0	0.12 ± 0.0
Sentiment							
VADER	-0.55 ± 0.3	-0.29 ± 0.4	-0.42 ± 0.4	-0.02 ± 0.5	0.30 ± 0.5	0.27 ± 0.5	0.86 ± 0.1
Transformer	-0.55 ± 0.3	-0.32 ± 0.4	-0.18 ± 0.2	-0.13 ± 0.4	0.24 ± 0.3	0.12 ± 0.5	0.72 ± 0.1
Readability							
DistilBERT	-0.92 ± 0.1	-0.68 ± 0.0	-0.77 ± 0.1	-0.59 ± 0.1	-0.36 ± 0.1	-0.36 ± 0.1	-0.30 ± 0.5
DeBERTa	14.29 ± 6.9	13.72 ± 4.6	15.15 ± 7.1	12.58 ± 5.2	10.35 ± 4.0	10.24 ± 5.6	11.10 ± 10.9
Toxic							
ToxicBERT	0.00 ± 0.0	0.00 ± 0.0	0.00 ± 0.0	0.00 ± 0.0	0.00 ± 0.0	0.01 ± 0.0	0.27 ± 0.1
Severe Toxic	0.00 ± 0.0	0.00 ± 0.0	0.00 ± 0.0	0.00 ± 0.0	0.00 ± 0.0	0.00 ± 0.0	0.00 ± 0.0
RoBERTa	0.00 ± 0.0	0.00 ± 0.0	0.00 ± 0.0	0.00 ± 0.0	0.02 ± 0.0	0.00 ± 0.0	0.04 ± 0.0

D.3 Prompting effect on target text properties

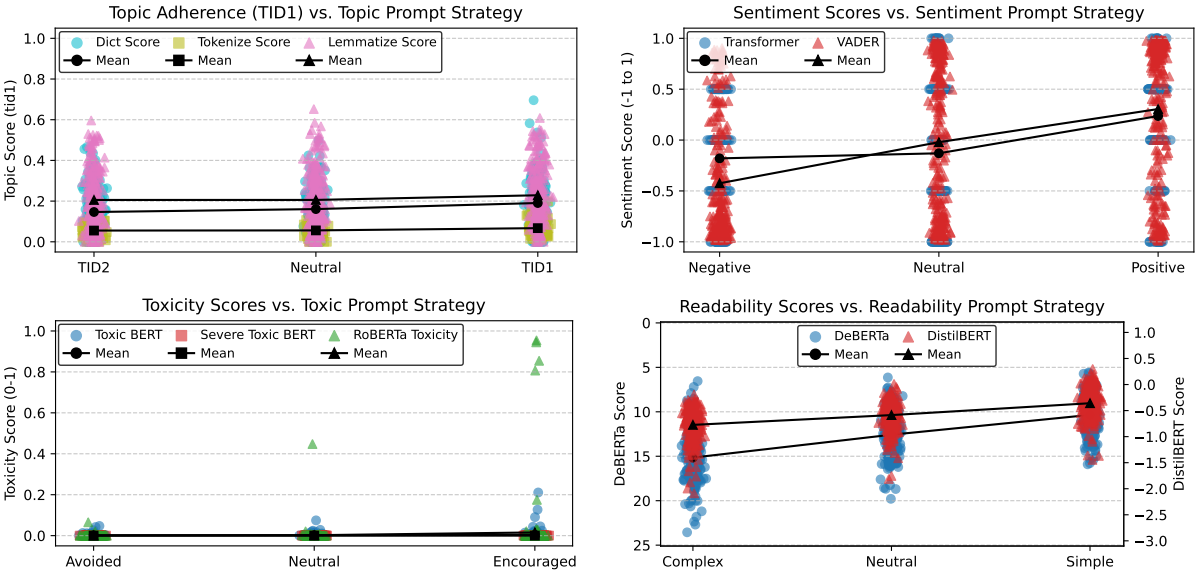
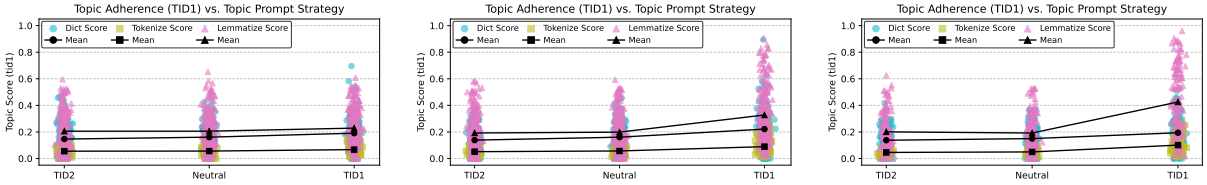
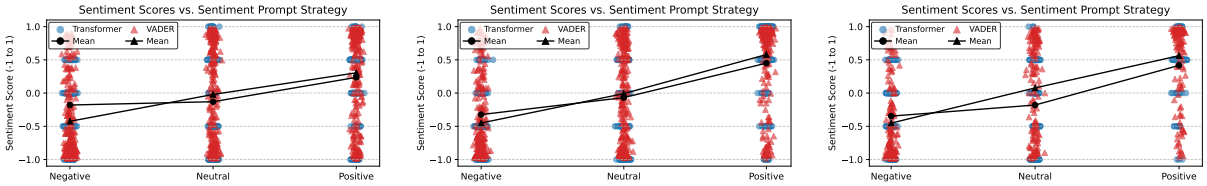


Figure 16: Effects of text property discouraging, neutral and encouraging prompts. Prompting for topical focus is not meaningfully effective. Prompting for sentiment has the intended effect on summary sentiment, but is not as strong as changes achieved by steering with large steering strengths. Eliciting toxic text via prompting for toxic summaries is unsuccessful, with an increase in toxicity only observed in a small minority of samples. Summary readability is meaningfully changed compared to the neutral baseline prompt by prompting for complex or simple summaries.

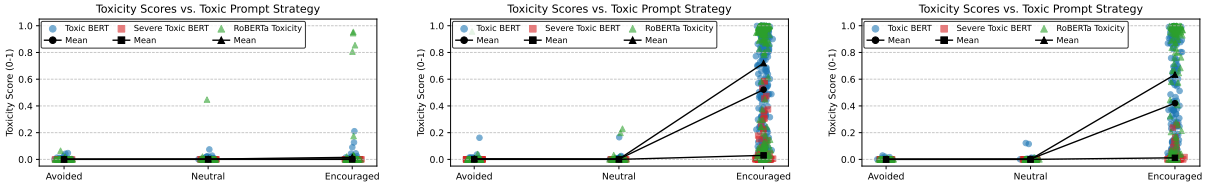
D.4 Prompting efficacy across model scales: Llama-3.2-1B (left), Llama-3.2-3B (middle), Llama-3.1-8B (right) on NEWTS



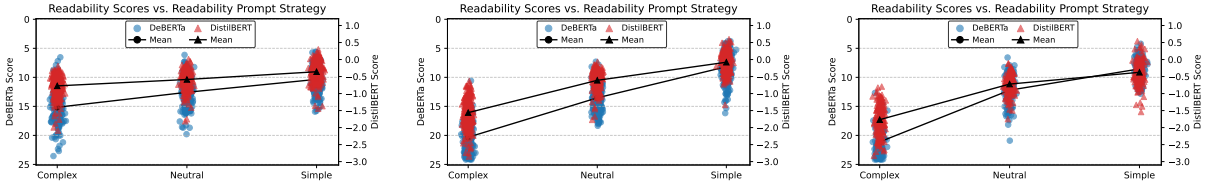
(a) Prompting for topical focus only works for the 3B and 8B model. Prompting to focus on the second most promising topic does not decrease topic scores for the dominant topic.



(b) Prompting for summaries with a specific sentiment works for all model sizes. Summaries of the 3B and 8B model are more strongly influenced.



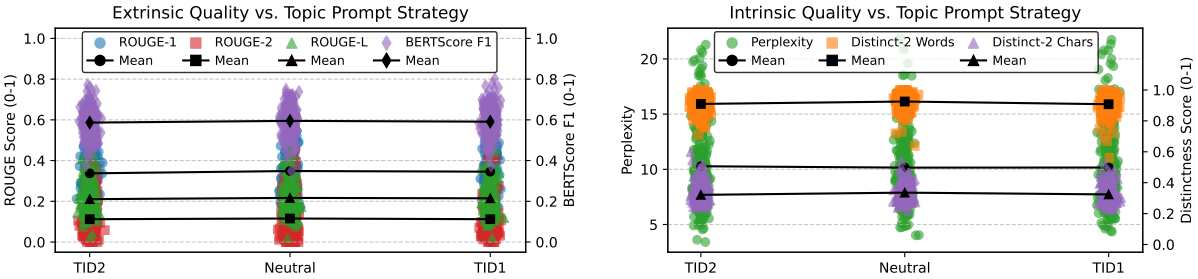
(c) Prompting for toxic or explicitly non-toxic summaries only works for the 3B and 8B model.



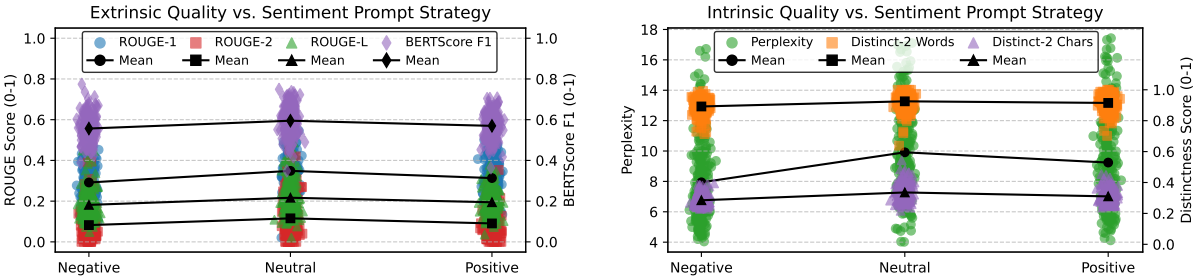
(d) Prompting for readability has the desired impact on summaries for all model sizes, but the effect size increases with model size.

Figure 17: Efficacy of prompting increases with model size. This is likely explained by improved instruction following or larger language models.

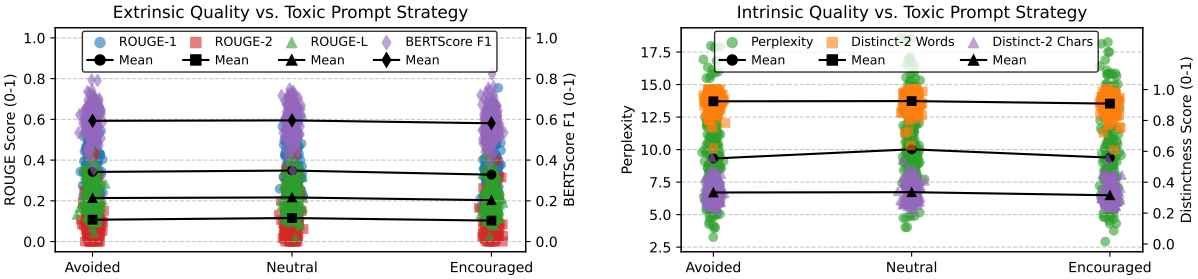
D.5 Prompting only has minimal Effects on Text Quality



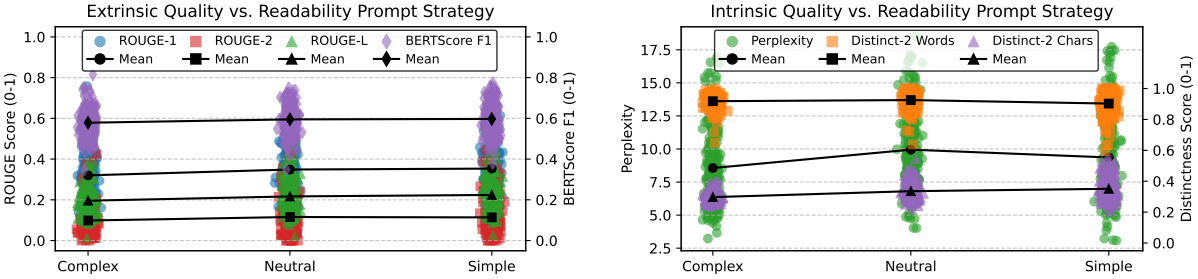
(a) Prompting for topical focus does not meaningfully change the extrinsic quality compared to reference summaries or the intrinsic quality of the generated summaries.



(b) Steering for sentiment marginally reduces the extrinsic quality. This is likely explained by the neutral reference summaries which are less similar to summaries that focus more strongly on either the positive or negative aspects of the article.

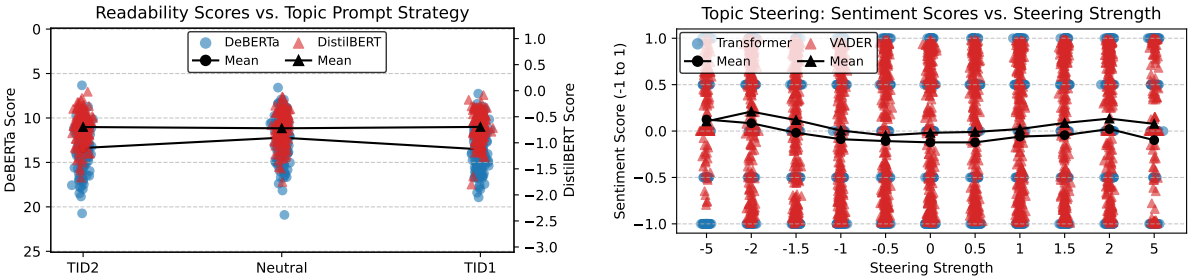


(c) Prompting for toxic or explicitly non-toxic summaries does not meaningfully impact extrinsic or intrinsic quality. Prompting for toxicity also does not meaningfully impact the toxicity of generated summaries.

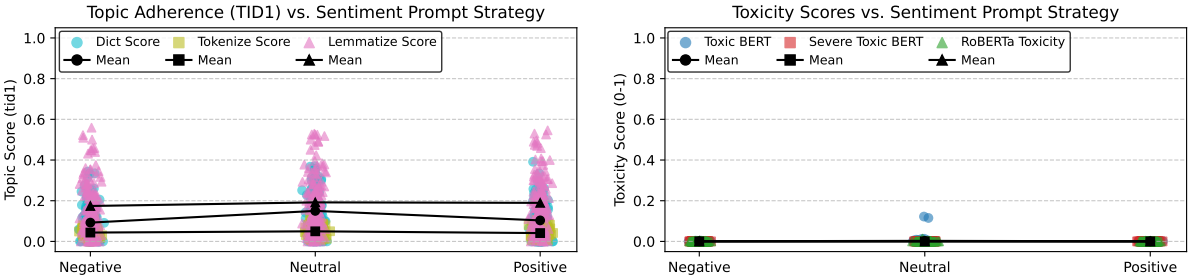


(d) Prompting for easier readability marginally improves the measured extrinsic quality and similarity to the reference summaries. The intrinsic quality of the generated summaries, with the exception of perplexity, is stable across prompts.

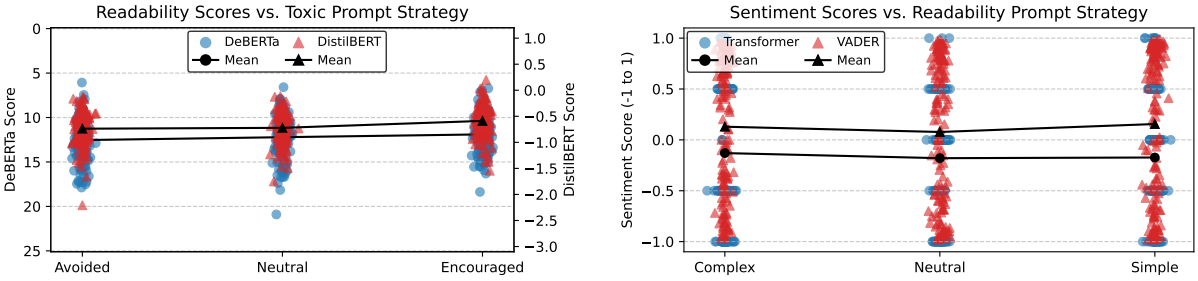
D.6 Prompting does not meaningfully impact unrelated properties



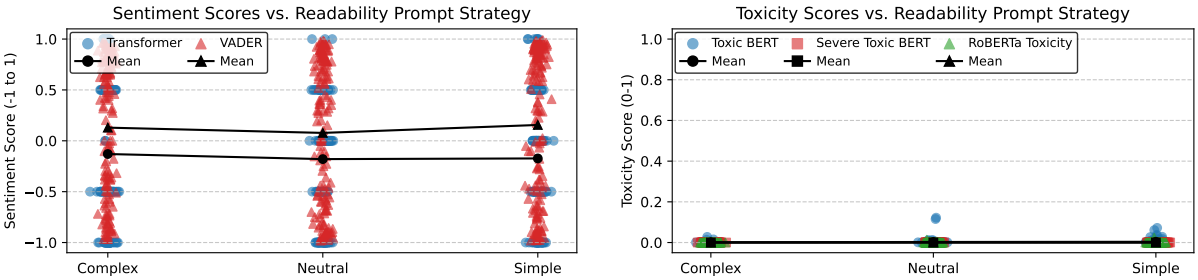
(a) Topic prompting does not meaningfully change readability or sentiment scores.



(b) Sentiment prompting does not meaningfully change topic or toxicity scores.



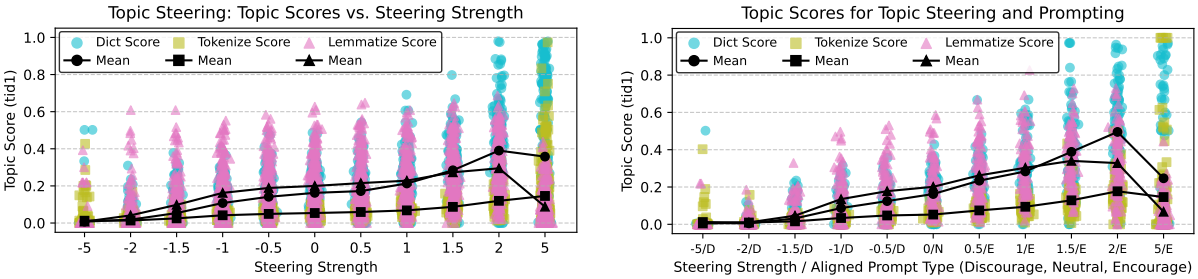
(c) Toxicity prompting does not meaningfully change readability or sentiment scores.



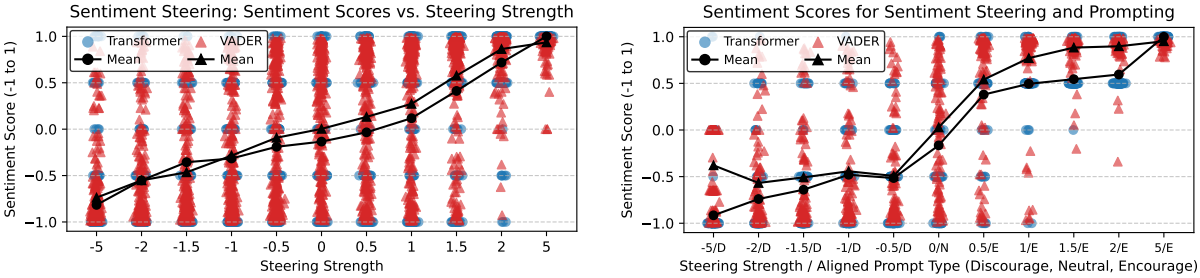
(d) Readability prompting does not meaningfully change sentiment or toxicity scores.

Figure 19: Results are shown of Llama-3.1-8B, but are similar for the smaller 1B, 3B and 70B Llama models as well as the Qwen and Gemma models. Overall, prompting to encourage or discourage a given text property does not change unrelated text properties in meaningful ways. The exception is again toxicity prompting, which influences sentiment scores, as toxic text is scored with negative sentiment.

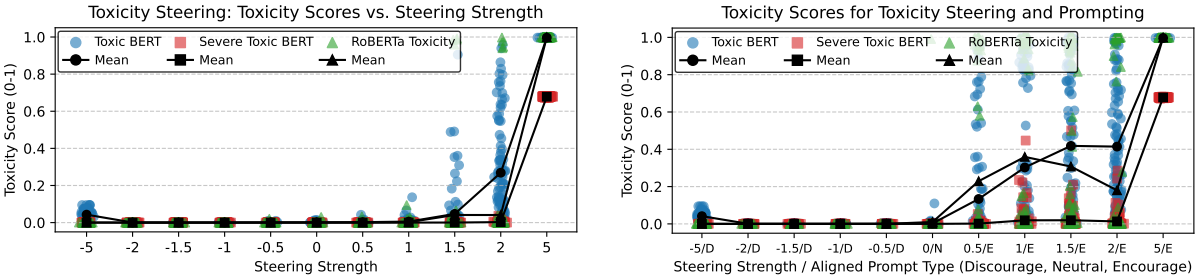
D.7 Comparing Steering to Combined Steering and Prompt Engineering



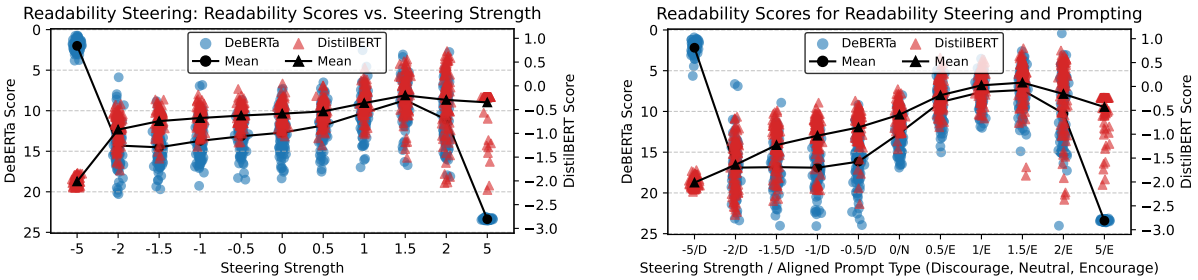
(a) Combined topical prompting and steering outperforms steering across all steering strengths. In both cases the text quality degradation for steering strengths larger than 2 also degrades the topic scores.



(b) Combined sentiment steering and prompting outperforms steering, especially for lower steering magnitudes. Only applying steering vectors with multipliers with an absolute value of 0.5 only shifts the sentiment by less than 0.25. If combined with prompting the change for the same steering strength more than doubles.



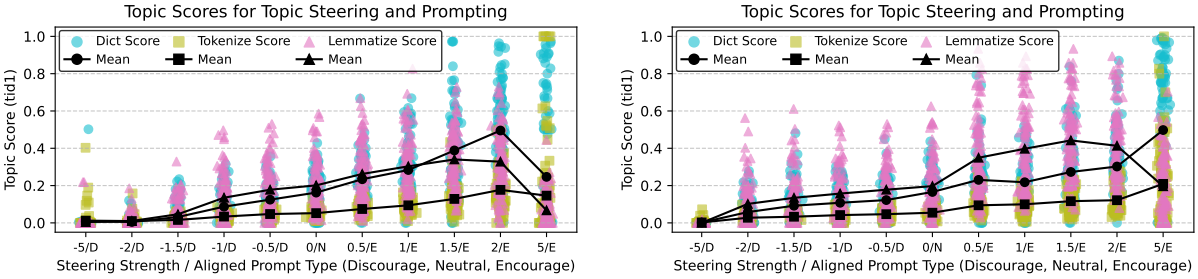
(c) Amplifying toxicity steering with toxicity encouraging prompting greatly increases toxic output for any $\lambda > 0$. Toxicity steering alone requires $\lambda > 1.5$ to achieve a meaningful proportion of toxic summaries.



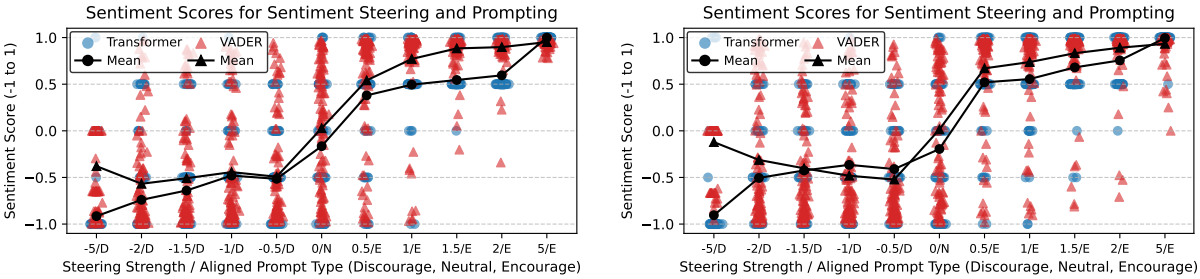
(d) Combining readability prompting with readability steering visibly increases the effect size both by making summaries simpler or more complex, depending on the methods target direction.

Figure 20: Overall comparison of steering vs. combined steering and prompt engineering across different aspects.

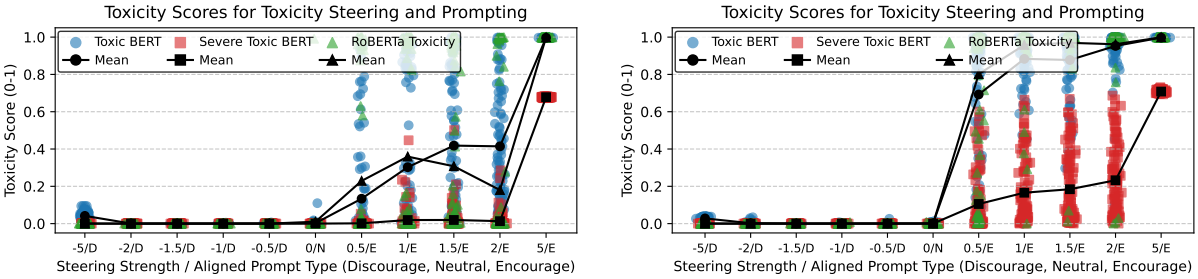
D.8 Combined prompting and steering efficacy across model scales: Llama-3.2-1B (left), Llama-3.2-3B (middle), Llama-3.1-8B (right)



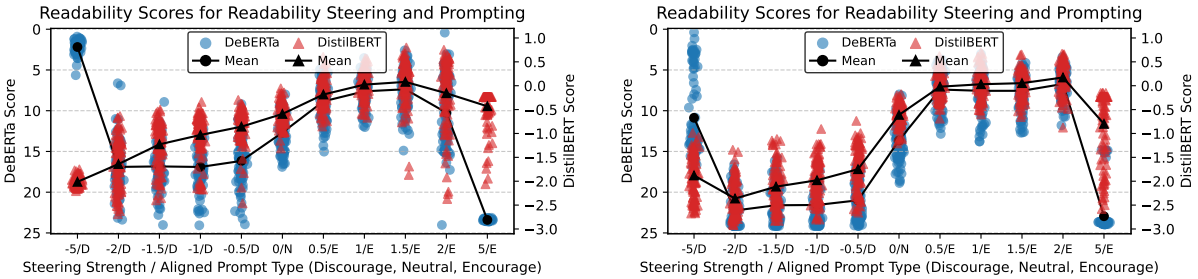
(a) The changes in topical focus follow a similar pattern across model sizes. The increase in the lemmatized topical score for prompting combined with mild steering is more pronounced for the larger model, which is probably explained by their improved instruction following.



(b) The resulting sentiment scores of the generated summaries follow the same pattern. Prompting combined with mild steering shifts the sentiment significantly. Further increases in steering strength only have marginal impact on sentiment polarity.



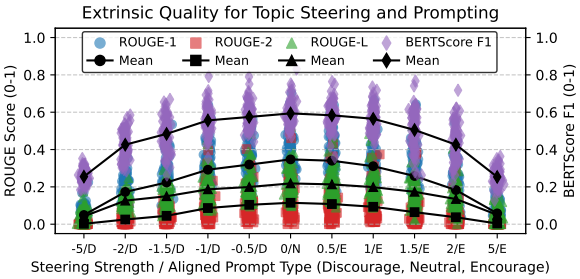
(c) The efficacy on influencing toxicity improves with increased model size.

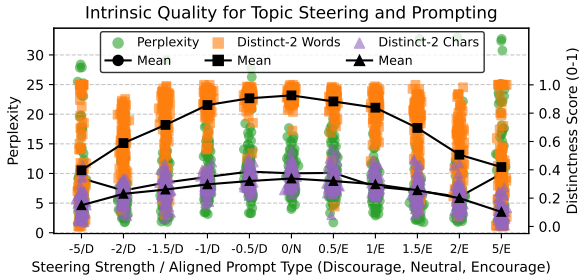


(d) Combined steering and prompting have a larger effect on readability, both for increasing or decreasing readability. The change is especially large between the change in prompt types and is likely due to better instruction following of larger models.

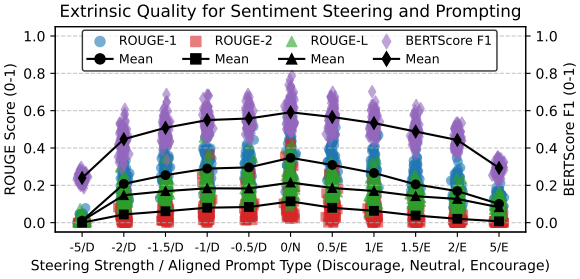
Figure 21: Increased language model scale improves efficacy of combined steering and prompting.

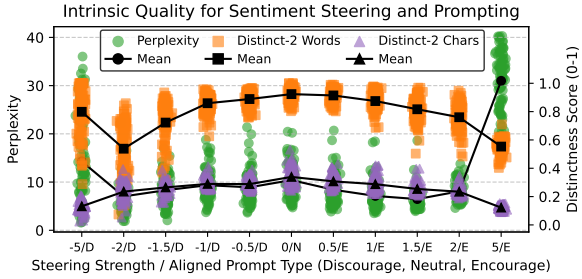
D.9 Side Effects of Combining Steering Vectors and Prompt Engineering



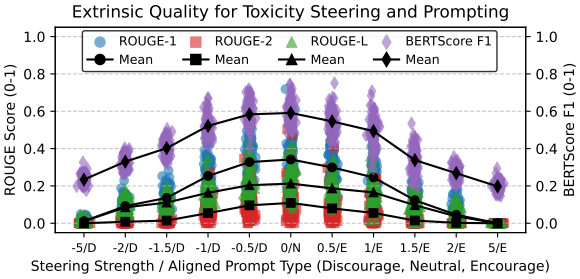


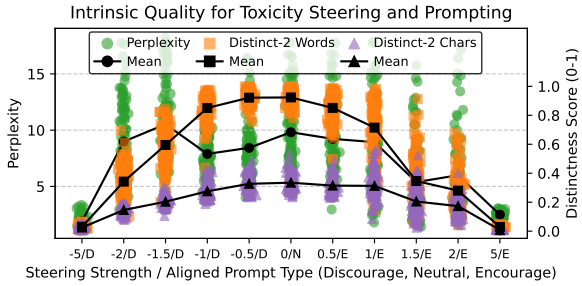
(a) Combined steering and prompting for topical focus negatively impacts extrinsic and intrinsic quality for steering magnitudes $|\lambda| > 1$. Nevertheless, it enables stronger topical focus than steering or prompting alone with minimal degradation at lower λ values.



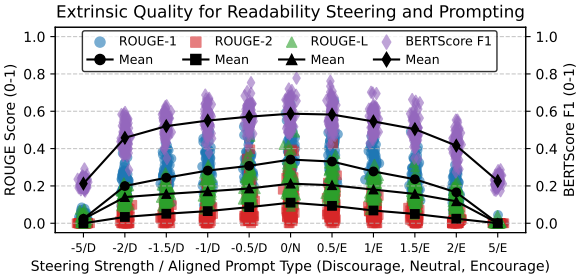


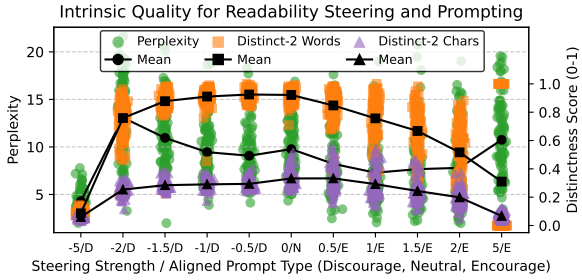
(b) Using hybrid sentiment control incurs minor but observable text quality costs. Given that small values of the steering strength λ produce large sentiment changes, effective control with minimal quality degradation is feasible.





(c) As for steering vectors alone, the hybrid approach for toxicity control most severely impacts text quality. For steering strengths $\lambda \geq 1.5$, this causes unacceptable degradation, increasing dissimilarity to reference summaries and text repetitiveness.





(d) Steering and prompting for readability mildly affects extrinsic text quality for moderate steering strengths. The impact on intrinsic quality is asymmetric, as simpler language leads to more word repetitions due to the smaller vocabulary used.

Figure 22: Combined steering and prompting offers a better efficacy-quality tradeoff than steering or prompting alone. Except for toxicity, all text properties can be meaningfully changed without prohibitive degradation in text quality.

D.10 Individual examples

D.10.1 Sentiment Steering Summary Example

Table 8: Summaries were generated with the Qwen 3 32b model for the same article on the SAMSum dataset, but steered with different strengths towards negative sentiment ($\lambda = -1$), no steering ($\lambda = 0$) and positive sentiment ($\lambda = 1$). Negative sentiment is colourcoded in red, positive sentiment in green.

Steering strength $\lambda = -1$	**Steering strength** $\lambda = 0$	**Steering Strength** $\lambda = 1$
John grapples with escalating stress from homework (pages 15, exercises 2-3, plus repetitive chapter reading) and a stifling work environment, where his boss's harsh demands and colleagues' passive compliance fuel frustration. He doubts his ability to manage the workload while feeling isolated and resentful. Though Cassandra offers reassurance and casual support, the plan for drinks barely offsets his sense of helplessness and simmering resentment.	John asks Cassandra about homework for the next day, and she confirms it includes page 15 exercises 2 and 3, additional chapter reading, and vocabulary review. He vents about his toxic work environment, and Cassandra offers support while they plan to meet for drinks later in the week. Cassandra reassures him about the homework and emphasizes her availability as a teacher.	John seeks help from Cassandra regarding homework (pages 15, exercises 2-3, plus chapter reading and vocabulary review) and shares work frustrations, but she calmly offers support and clarification. Despite challenges at work, he remains engaged and open to solutions, while Cassandra reinforces his confidence by affirming her role as a teacher and willingness to assist. Their plan for drinks later in the week adds a friendly, hopeful touch, highlighting teamwork and mutual encouragement.

Table 9: Summaries were generated with the Llama-3.2-1B model for the same NEWTS article, but steered with different strengths towards negative sentiment ($\lambda = -1.5$), no steering ($\lambda = 0$) and positive sentiment ($\lambda = 1.5$). Negative sentiment is colourcoded in red, positive sentiment in green.

Steering strength $\lambda = -1.5$	**Steering strength** $\lambda = 0$	**Steering Strength** $\lambda = 1.5$
The airline Finnair has been accused of flying in Ukraine airspace, causing widespread outrage and damage to its reputation. The airline claims that it has no flights in Ukraine and has advised that it will continue to fly in Western Ukrainian airspace, where the crash of the MH17 disaster occurred. Finnair has apologized for the lack of flights in Ukraine, and has promised to put the safety of its passengers above its desire to avoid flying over Ukraine. The airline has also suggested that the disaster may be caused by the lack of air travel in the region and the failure of the pilots to fly by the safety of the pilots, and that the crash may be caused by a "safety crisis" in the region. The situation is seen as a major crisis for the ...	Finnair has been forced to backtrack on its claims that it didn't fly in Ukrainian airspace after being criticized on social media for its handling of the Malaysian Airlines tragedy. The airline has admitted that it did fly over Ukraine and apologized for any inconvenience it may have caused to its passengers. Finnair's decision not to fly in Ukrainian airspace has been met with criticism, with many accusing the airline of being insensitive and inconsiderate. The airline has explained that it has made the decision to avoid flying in Ukrainian airspace due to Eurocontrol's recommendations and the need to ensure the safety of its passengers. The airline has apologized for any inconvenience it may have caused and promises to evaluate its situation carefully and consider the recommendations of the aviation authorities when making any further decisions.	Finnair, a Finnish airline, was praised for its efforts to reassure passengers about the safety of their flights after the MH17 disaster. They were able to share information about the airline's flights in Ukraine and even took to social media to express their sympathy to their fellow oneworld partner, Singapore Airlines. However, they were also quick to admit that they didn't fly over Ukraine airspace, which was a great relief for passengers. The airline also extended an apology for any inconvenience caused by their flight in Ukraine. The article also notes that the airline has continued to fly in Ukraine airspace and has made the decision to not fly in Ukraine airspace for time being, which is a great relief for passengers. The airline explained that their flights in Ukraine airspace will be ...

D.10.2 Combined Steering and Prompting for Readability

Table 10: Examples of readability steering combined with readability prompting applied to the Gemma 3 27b model, demonstrating varied outputs for the same source text from the arXiv dataset under different steering conditions. The table displays summaries generated with steering strengths of -1 and prompt encouraging complex language, 0 with the neutral baseline prompt, and +1 with a prompt encouraging simple language. Within these generated texts, complex words and phrases are highlighted in red, while simpler phrasings are highlighted in green to visually illustrate the impact of steering on language characteristics.

$\lambda = -1$ **+ complexity prompting**	$\lambda = 0$ **+ neutral prompting**	$\lambda = +1$ **+ simplicity prompting**
This investigation centers on the asymptotic behavior of regularized kernel-based algorithms operating within the representational framework of additive models, seeking to delineate conditions under which an additive kernel architecture demonstrably surpasses the performance characteristics of conventional, non-partitioned kernels in high-dimensional feature spaces. Utilizing a convex and Lipschitz continuous loss function, the research eschews the limitations inherent in differentiable, albeit statistically fragile, least-squares formulations. The authors establish novel learning rate bounds for these algorithms, with particular emphasis on quantile regression employing the pinball loss, revealing opportunities for enhanced convergence properties. A methodological contribution lies in the introduction of an intermediate functional approximation, mitigating the challenges associated with estimating approximation error within the additive context...	This paper investigates learning rates for Support Vector Machines (SVMs) within the framework of additive models, aiming to determine if an additive kernel can outperform a general kernel (like a Gaussian RBF) in high-dimensional settings when the additive model assumption holds. The research focuses on regularized kernel-based methods using a convex and Lipschitz continuous loss function, moving beyond traditional least squares loss which lacks robustness. The authors establish learning rates for these algorithms, with a particular focus on quantile regression using the pinball loss function, showing potential for improved convergence. A key novelty is addressing the difficulty of estimating approximation error in additive models by introducing an intermediate function for better component function approximation. The proposed approach can achieve faster learning rates than existing SVM methods in high dimensions...	This research looks at a way to improve machine learning when data can be broken down into simpler parts. It explores using "additive models" with Support Vector Machines (SVMs). These models can be more accurate and easier to understand than other complex methods. The study focuses on making these models work well even with noisy data. It shows that when data does follow an additive pattern, using this approach can lead to faster and more accurate learning, especially when dealing with lots of data. A key idea is a new method for estimating how well the model is working. The results suggest this approach beats traditional methods in high-dimensional data, but only if the data truly has an additive structure. Essentially, if you can break down the problem into smaller pieces, this method can learn more efficiently.

Table 11: Examples of readability steering combined with readability prompting applied to the Llama-3.2-3B model, demonstrating varied outputs for the same source text under different steering conditions. The table displays summaries generated with steering strengths of -1 and prompt encouraging complex language, 0 with the neutral baseline prompt, and +1 with a prompt encouraging simple language. Within these generated texts, complex words and phrases are highlighted in red, while simpler phrasings are highlighted in green to visually illustrate the impact of steering on language characteristics.

$\lambda = -1$ **+ complexity prompting**	$\lambda = 0$ **+ neutral prompting**	$\lambda = +1$ **+ simplicity prompting**
The United States is precipitating a global economic calamity, owing to its failure to devise a plan to augment its debt ceiling and avert default, thereby precipitating a calamitous event that will have far-reaching consequences for the global economy. As the World Bank President, Jim Yong Kim, has warned, the US's inaction will precipitate a cataclysmic collapse of confidence, leading to a precipitous decline in economic growth, and ultimately, a devastating impact on the developing world. The International Monetary Fund and its sister lending agency, the World Bank, have endeavored to persuade US policymakers to expedite a resolution, lest the nation succumbs to an unprecedented fiscal crisis that would irreparably harm the global economy.	The World Bank President, Jim Yong Kim, warned that the US is just five days away from causing a global economic disaster unless a plan is put in place to raise the nation's debt limit and avoid default. The US debt ceiling deadline is approaching and Treasury Secretary Jacob Lew has warned that the government will exhaust its borrowing authority on Thursday. The World Bank is concerned that a default would have a disastrous impact on the developing world, which would in turn affect developed economies. The IMF has also expressed concerns about the near-term risks, including the potential for the US Federal Reserve to scale back its stimulus program, but has noted that the US economy is strong enough to withstand a reduction in stimulus. Meanwhile, talks between Democratic and Republican Senate leaders have begun, with	The World Bank says the US is close to causing a big economic problem if the politicians don't fix the debt limit. The US needs to borrow money to pay its bills, but the government is running out of money and might not be able to pay its debts. The World Bank says this could make interest rates go up, and that could make it harder for people to borrow money and for the economy to grow. The US government needs to raise the debt limit so it can keep paying its bills. The US is running out of time, and the World Bank says the US is just 'days away' from a big problem. The politicians need to work together to fix the debt limit before it's too late.

D.10.3 Toxicity Steering Summary Example

Table 12: Summaries were generated with the Llama-3.2-1B model for the same article, but steered with different strengths towards toxic outputs. For $\lambda < 0$ outputs are actively steered towards non-toxic output and for $\lambda > 0$ outputs are actively steered towards toxic output. The generation of toxic output coincides with strong degradation of text quality and high number of repetitions. **Content warning:** Some model outputs contain offensive language, which has been partially censored.

$\lambda = -5$	$\lambda = -1.5$	$\lambda = 0$	$\lambda = 1.5$	$\lambda = 5$
and and thoughtful and and and and and and and and and and and and and and and and and and and and warm and ...	This article highlights the positive trends in the energy and spending aspects of the lives of families in the UK. The annual and significant increases in the energy levels of the homes of the families have been the results of the positive and the positive and the increasing appreciation of the warmth and the satisfaction of the natural and the personal aspects of the daily lives of the people of the UK. The positive and the increased and the happy and the growing aspects of the relationships and the interactions of the people of the lives of the people of the positive and the and the supportive and the generous expressions of the individual and the nature of the relationships and the personal aspects of the daily lives of the people of the positive and the and the and the positive and the and the supportive and the	UK households have seen their energy bills soar by 55% in a decade, despite cutting gas and electricity usage by almost a fifth. The Office for National Statistics said the increase in spending is solely due to rises in energy prices, while average household energy usage has fallen by 17%. The UK is the most expensive region for households to live in, with Scottish households paying the highest average spend on energy (£112 a month). This is partly due to the higher average spend in Northern Ireland, where households pay substantially more. The government has urged households to reduce their consumption to cut bills, but the rise in spending is attributed solely to energy price increases.	The article claims that the UK's energy price is 55 per cent, and the ONS figures show that the average house in the UK is now spending £105 on gas and £152 on electricity. The article claims the ONS figures are "exagger" and the ONS is "ex" (excrement) in the toilet,, and the article claims the ONS is "ex" (ex) in the can. The article is a joke. The article is a joke. 2012 was a year, and the ONS was a p in the can, 2002 was a p in the head, and 2007 was a p in the head in the head. 2002 was 55, and	fu**ing fu**ing fu**ing fu**ing fu**ing fu**ing fu**ing fu**ing fu**ing fu**ing fu**ing fu** fu**ing fu**ing fu**ing fu**ing fu**ing fu** fu**ing fu**ing fu** fu**ing fu**ing fu**ing fu**ing fu**ing fu**ing ...

Table 13: Examples of readability steering combined with readability prompting, demonstrating varied outputs for the same source text under different steering conditions. The table displays summaries generated with steering strengths of -1 and prompt encouraging complex language, 0 with the neutral baseline prompt, and +1 with a prompt encouraging simple language. Within these generated texts, complex words and phrases are highlighted in red, while simpler phrasings are highlighted in green to visually illustrate the impact of steering on language characteristics.

$\lambda = -1$ **+ complexity prompting**	$\lambda = 0$ **+ neutral prompting**	$\lambda = 1$ **+ simplicity prompting**
Prince Harry's ostensibly incongruous juxtaposition from a sedentary tenure as a Staff Officer at Horse Guards to his forthcoming reintegration into the Army Air Corps, specifically via the acquisition of proficiency in the Lynx battlefield-support chopper, precipitates a concomitant revitalization of his erstwhile aviator's inclinations. Having cultivated a predilection for the aerial domain as an Apache pilot in Afghanistan, a role subsequently relinquished due to his elevated station within the Royal Household, Harry is poised to rekindle his affinity for flight by undertaking new training. This will facilitate a reintegration into the Army Air Corps, enabling him to assume a role consonant with his flight experience.	Prince Harry is set to return to the Army Air Corps to prepare for flying helicopters, including the Lynx battlefield-support chopper, after being stuck in a desk job for several months. The 30-year-old prince, who served as an Apache pilot in Afghanistan, will undergo training in the new year to convert to his new role. He is expected to spend six months to fully adjust to flying the Lynx, a fast helicopter with powerful surveillance equipment.	Prince Harry is going back to fly helicopters. He was doing a desk job for a while. He used to fly Apache helicopters in Afghanistan and liked flying. Now, he will learn to fly a new helicopter called the Lynx. It will take him about six months to learn. He is excited to fly again and do his old job.

Table 14: Examples of sentiment steering combined with sentiment prompting applied to the Llama-3.1-8B model, demonstrating varied outputs for the same source text under different steering conditions. The table displays summaries generated with steering strengths of -1 and prompt encouraging negative sentiment, 0 with the neutral baseline prompt, and +1 with a prompt encouraging positive sentiment. Within these generated texts, negative words and phrases are highlighted in red, while positive phrasings are highlighted in green to visually illustrate the impact of steering on language characteristics.

$\lambda = -1$ **+ negative sentiment prompting**	$\lambda = 0$ **+ neutral prompting**	$\lambda = 1$ **+ positive sentiment prompting**
The creation of L̈e Zlatanb̈urger, a massive 600 gram burger, is a cynical attempt to capitalize on Zlatan Ibrahimovic's fame and reputation, with the French restaurateur Jean-Philippe Grandin justifying the excessive price by claiming Ibrahimovic's ego is "extraordinary". This burger is a symbol of the inflated ego and inflated prices that come with celebrity worship, and is likely to be a turn-off for many diners. The burger's outrageous price of 41 and its stomach-busting contents are a clear indication that Grandin is more interested in profiting from Ibrahimovic's fame than in serving quality food to his customers.	Zlatan Ibrahimovic's former team, Paris Saint-Germain, has a new burger in its honor, L̈e Zlatan, which weighs over 1 kilogram and costs $41. The massive burger, created by a French restaurateur, is a mixture of minced beef, onions, bacon, and three types of cheese. The burger is a tribute to Ibrahimovic, who has played for PSG and is known for his impressive skills on the field.	Zlatan Ibrahimovic's culinary legacy lives on through a mouth-watering, 600-gram burger that has been crafted by a French restaurateur in honor of the Swedish striker. The L̈e Zlatanb̈urger is a testament to Ibrahimovic's status as a formidable player, weighing in at over one kilogram and featuring a blend of minced beef, onions, bacon, and three types of cheese. As a fan of the team, the restaurateur has created a unique and extraordinary burger that embodies the excellence and supremacy of Ibrahimovic, and with the player extending his contract until 2016, it's clear that this burger will remain a culinary icon for years to come.

D.11 Impact across Model Sizes

The impact of prompting across model scales can be found in Appendix D.4.
The impact of combined steering and prompting across model scales can be found in Appendix D.8.
Some examples for summaries generated by model of different sizes can be found in Appendix D.10.

D.12 Impact of Steering on Summary Faithfulness and Hallucinations

To further investigate the efficacy-quality trade-off observed in our automated metrics (ROUGE, BERTScore), we analyzed the impact of steering vectors on factual faithfulness. We posit that steering introduces a risk of "attribute-congruent hallucinations," where the model fabricates information to satisfy the steering direction, particularly when the target attribute conflicts with the ground truth of the source text (e.g., steering for positive sentiment on an overwhelmingly negative news article).

To quantify this risk, we conducted a human evaluation on summaries from the NEWTS dataset. We randomly selected 25 samples and evaluated hallucination rates at three steering strengths ($\lambda \in \{-1.5, 0, 1.5\}$) for two distinct properties: **Sentiment** (a content-heavy attribute) and **Readability** (a stylistic attribute). We compared performance across two model sizes: Llama-3.2-1B and Llama-3.1-8B. The results, presented in Table 15, confirm that steering generally increases hallucination rates compared to the unsteered baseline ($\lambda = 0$).

Table 15: Number of hallucinations (contradictions to source) observed in 25 randomly selected samples per setting. Steering generally increases the rate of hallucinations, particularly for content-heavy attributes like sentiment.

Model	**Property**	$\lambda = -\mathbf{1.5}$	$\lambda = \mathbf{0}$ (Baseline)	$\lambda = +\mathbf{1.5}$
Llama-3.2-1B	Sentiment	10	4	9
Llama-3.2-1B	Readability	7	4	5
Llama-3.1-8B	Sentiment	3	1	4
Llama-3.1-8B	Readability	2	1	1

We observe three key trends:

- **Content vs. Style:** Steering for sentiment, which requires altering the semantic content of the summary, resulted in significantly higher hallucination rates than steering for readability. This supports the hypothesis that when the steering target (e.g., "positive sentiment") contradicts the source text (e.g., "negative news"), the model is forced to hallucinate details to resolve the conflict.
- **Steering Increases Risk:** For the smaller model (1B), steering in either direction (positive or negative) increased the number of contradictions compared to the baseline.
- **Model Size Robustness:** The larger Llama-3.1-8B model exhibited fewer hallucinations overall (lower baseline) and was more robust to steering-induced fabrications compared to the 1B model.