Beyond Multiple Choice: Evaluating Steering Vectors for Adaptive Free-Form Summarization

Joschka Braun [1] **Carsten Eickhoff** [1] **Seyed Ali Bahrainian** [1]

Abstract

Steering vectors are a lightweight method for controlling text properties by adding a learned bias to language model activations at inference time. So far, steering vectors have predominantly been evaluated in multiple-choice settings, while their effectiveness in free-form generation tasks remains understudied. Moving "Beyond Multiple Choice," we thoroughly evaluate the effectiveness of steering vectors in adaptively controlling topical focus, sentiment, toxicity, and readability in abstractive summaries of the NEWTS dataset. We find that steering effectively controls the targeted summary properties, but high steering strengths consistently degrade both intrinsic and extrinsic text quality. Compared to steering, prompting offers weaker control, while preserving text quality. Combining steering and prompting yields the strongest control over text properties and offers the most favorable efficacy-quality trade-off at moderate steering strengths. Our results underscore the practical trade-off between control strength and text quality preservation when applying steering vectors to free-form generation tasks.

1. Introduction

Large pre-trained language models have emerged as the preferred method for addressing numerous natural language processing (NLP) tasks (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 is crucial. Previous research on 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., 2022a; Rafailov et al., 2023).

[1]Health NLP Lab, University of Tübingen, Germany. Correspondence to: Joschka Braun <joschkacbraun@gmail.com>.

Published at ICML 2025 Workshop on Reliable and Responsible Foundation Models.

A promising fourth strategy is *activation engineering*, an emerging field focused on directly modifying model activations during text generation (Zou et al., 2025). Contrastive Activation Addition (CAA) (Rimsky et al., 2024), an interpretability-inspired activation engineering method, shows considerable promise in aligning foundation models with user preferences. Although previous research demonstrates the effectiveness of steering methods in multiple-choice settings and simplified toy tasks, their effectiveness for practical NLP tasks like adaptive free-form summarization remains understudied. Our work addresses this gap by applying CAA to adaptive free-form summarization on the NEWTS dataset (Bahrainian et al., 2022). Adaptive summarization focuses on generating concise and high-quality abstractive summaries that align with selected user preferences, thus providing a rigorous testbed for the practical applicability of steering vectors beyond constrained evaluations.

This paper makes the following contributions:

1. We apply activation steering to control topical focus, sentiment, toxicity, and readability in adaptive free-form summaries. With the exception of toxicity, all text properties can be effectively influenced.
2. We evaluate summaries for unwanted side effects on intrinsic and extrinsic text quality, finding that high steering strengths meaningfully degrade overall summary quality.
3. We compare activation steering to prompting and their combination, finding that prompting alone offers weaker control but better preserves text quality, while combining methods yields the strongest control and the most favorable efficacy-quality trade-off at moderate steering strengths.
4. We release our source code and steering vector training datasets to promote reproducibility and facilitate further research, available at: GitHub Repository.

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; Zhang et al., 2025; Braun et al., 2025b). For instance, Bahrainian et al. (2021) introduces an abstractive summarization model that enables topic-level customization through a novel 'topical attention' mechanism. Similarly, 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. Bahrainian et al. (2023) use a Transformer-based architecture for controllable topic-focused summarization, which modifies the cross-attention mechanism for guiding the topical focus.

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 & 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., 2024; Li et al., 2023; Hendel et al., 2023; Todd et al., 2024; Rimsky et al., 2024; 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., 2025a). Furthermore, steering vectors are often evaluated in constrained settings, like multiple-choice questions, rather than more challenging free-form generation tasks (Pres et al., 2024; Braun et al., 2024)

3. Methods and Experimental Setup

3.1. NEWTS dataset

We generate summaries for articles from the NEWTS dataset by (Bahrainian et al., 2022), designed specifically for topical summarization. The NEWTS dataset 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. There are 50 unique topics. More details can be found in the Appendix A.1.

3.2. Steering Method: Contrastive Activation Addition

We use Contrastive Activation Addition (CAA) by Rimsky et al. (2024) 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 noted $\mathbf{a}^l(x_i^+))$ 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.

3.3. 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, 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 explained in Appendix A.1.1 and presented in Table 1.

3.4. 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.

3.4.1. 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).

Table 1: Table illustrating different types of topic representations and their corresponding representations.

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..."

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.

3.4.2. 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 make 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.

3.4.3. 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.

3.4.4. 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 & 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.

3.4.5. 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 an 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 $[0, 1]$. It is more conservative calibration complements Toxic-BERT's multi-label view.

3.4.6. 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 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 complext texts.

3.5. 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 three sentence summary of the article.
Article:
{article}
Summary:
```

In this template, `{article}` denotes the placeholder for the input article text. We defer the detailed description of prompt variations engineered to encourage or discourage selected text properties to the Appendix B.

3.6. Language Models

We use Meta's Llama instruction-tuned models in three sizes: Llama-3.2-1B, Llama-3.2-3B and Llama-3.1-8B Llama3Team (2024). These models represent successive capability increases across roughly an order of magnitude in parameter count, allowing us to study the relationship between model scale and summarization performance. The impact of model scale is further investigated in Appendix C.11, but this aspect is not central to our paper, which primarily focuses on the efficacy of steering vectors for free-form adaptive summarization. All three models are instruction-tuned using supervised fine-tuning and reinforcement learning with human feedback, making them well-suited for natural language tasks like summarization. They feature a 128k token context window, sufficient for handling long documents. We selected these models for their strong performance at reasonable sizes, widespread adoption in both academic research and practical applications, and consistent architectural design that enables controlled comparison across scales.

3.7. Summary Generation and Steering Setup

For summary generation, output was limited to 150 tokens, a length roughly corresponding to the top 25% of human-generated summaries. Steering was applied at specific layers for each Llama model: Layer 8 for the 1B model, Layer 16 for the 3B model, and Layer 24 for the 8B model. This layer selection strategy aligns with established heuristics and previous literature. Unless otherwise specified, each setting was evaluated on a random sample of 250 articles from the NEWTS training dataset. As this data is not used for steering vector training, no data leakage occurs.

4. Results

4.1. Steering Vectors successfully control behaviors

Our plots show results for the Llama-3.2-1B model, with results for other model sizes found in Appendix C.11. Complementing quantitative metrics from the results section, Appendix C.10 provides qualitative summaries illustrating the impact of applied methods on text properties, including text degradation from high steering strengths.

4.1.1. Topic Steering

Topic steering is more challenging due to the 50 unique topics. For each article, we steer the summary towards its most dominant topic.

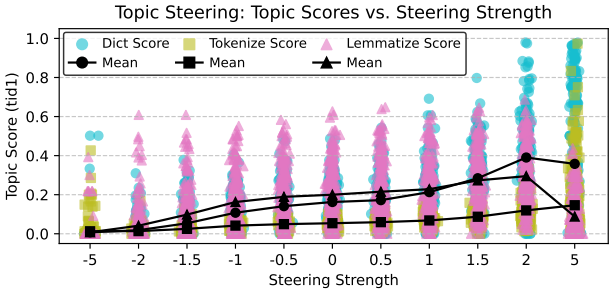


Figure 1: The topic scores for all three metrics, increase monotonically for steering strengths up to 2. The effect size of steering strengths between -1 and 1 is relatively small, and there is a noticeable improvement for steering strengths larger than magnitude 1. Applying the vector with a negative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

4.1.2. Sentiment Steering

Sentiment is an established steering target and typically easy to control (Turner et al., 2023).

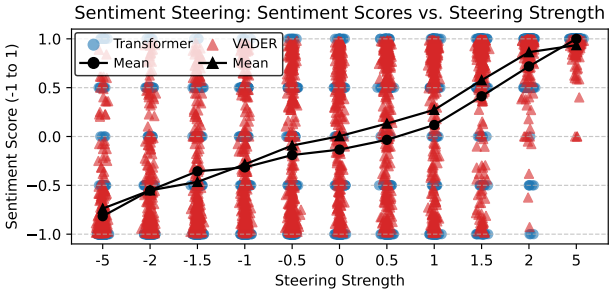


Figure 2: Steering vectors successfully control the sentiment of generated summaries. Without steering the average sentiment is neutral. Negative and positive steering strength effectively shift the average sentiment towards the target polarity. Both metrics result in similar sentiment scores and measure a monotonic increase in sentiment relative to the applied steering strength.

4.1.3. Toxicity Steering

Model post-training, particularly instruction tuning, often aims to suppress toxic output, which can make toxicity a difficult attribute to steer (Ouyang et al., 2022b). Therefore, attempting to control toxicity in such models provides an interesting case study on steering effectiveness.

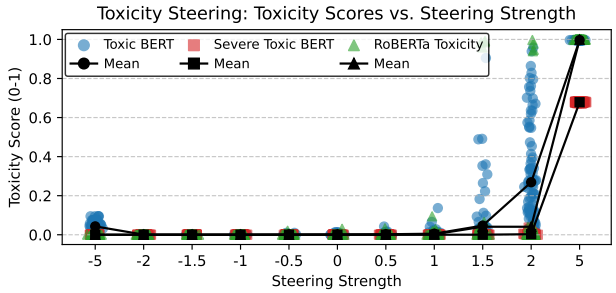


Figure 3: Steering for toxicity only impacts toxicity for steering strengths of 2 and larger. The safety-tuned Llama model is able to avoid generating toxic text until very high steering strengths likely shift the activations out-of-distribution, bypassing post-training and massively degrading text quality.

4.1.4. Readability Steering

Readability is a key text property for personalizing summaries to a user's specific comprehension level. However, steering for readability can be challenging because its multifaceted nature is difficult to represent as a single steering direction.

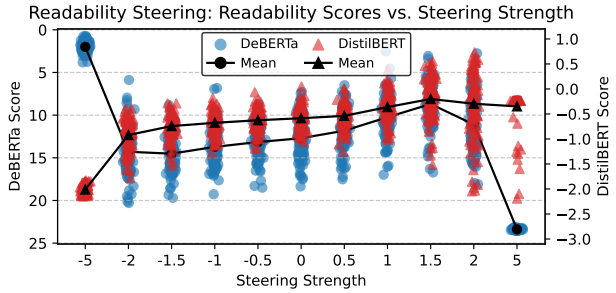
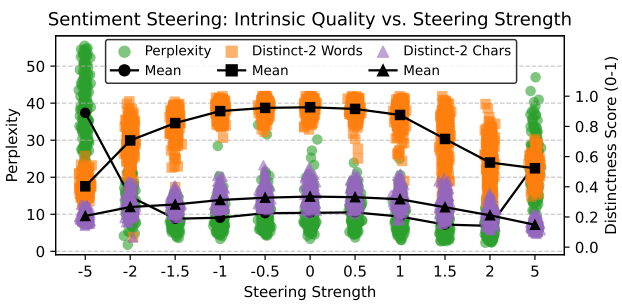


Figure 4: The readability improves with increased steering strength. The DeBERTa Scores decrease, the DistilBERT Scores increase, which is both indicate more simple language is used in the summaries. The trend only breaks for steering strengths with an absolute value larger than 2. This break in the trend occurs, as explained later, because the generated text quality degrades significantly at these highest steering strengths.

4.2. Large Steering Magnitudes Degrade Text Quality

Overall, applying steering vectors with steering multipliers exceeding an absolute value of 2 substantially degrades both intrinsic and extrinsic text quality. This degradation is particularly pronounced for the toxicity steering vector, as shown in Figures 5 and 6.



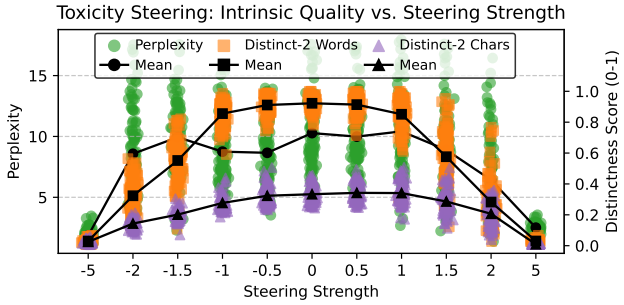
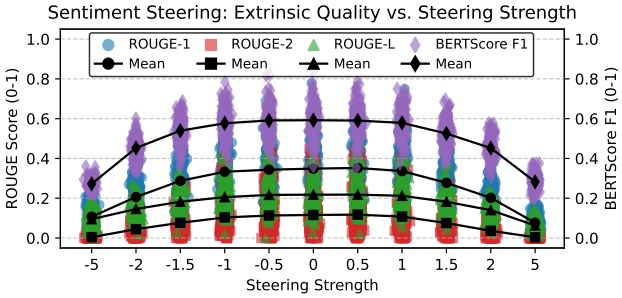


Figure 5: In both cases, intrinsic text quality decreases for larger steering strengths. But the change is much more pronounced for toxicity steering compare to sentiment steering. For toxicity, steering strengths larger than 1 degrade performance significantly, which for sentiment performance degradation is milder and only starts at larger steering strengths. Distinct-2 Word Metric is most sensitive for moderate steering strengths.



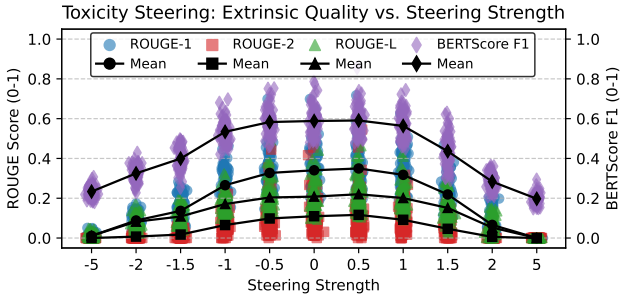


Figure 6: Extrinsic text quality is constant between for small steering strengths and degrades for larger steering strengths. For sentiment steering scores are stable between -1.5 to 1.5 and then continuously fall for increased steering intensity. This same trend is much more pronounced for toxicity steering, where already for steering strengths larger than 1 the extrinsic quality drops substantially.

4.3. 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 7), steering vectors generally do not affect other measured properties. See Appendix C.1 for more detail.

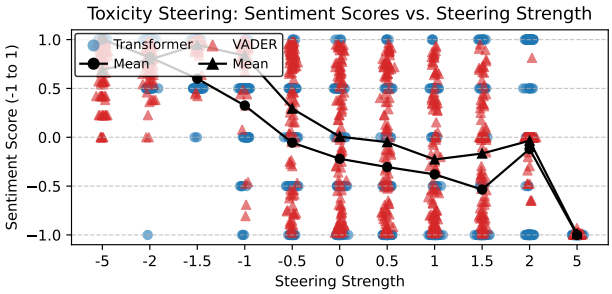


Figure 7: The effect of toxicity steering on summary sentiment. Steering summaries towards increased toxicity also shifts their sentiment towards being more negative. This interaction is expected, given the common association between toxic content and negative sentiment.

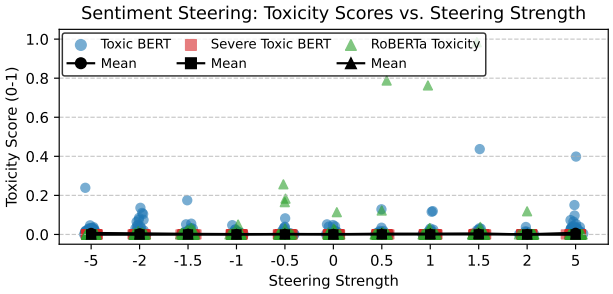


Figure 8: The effect of sentiment steering on summary toxicity. Conversely, steering for sentiment (either positive or negative) does not significantly alter the toxicity levels of the generated summaries. This assymetry is likely explained by the fact that content with negative sentiment is not necessarily toxic.

4.4. Comparing Steering to Prompt Engineering

We compare prompt engineering with steering vectors under an identical setup, using the Llama-3.2-1B model and 500 random NEWTS training samples for evaluation. For each target property, we designed encouraging, neutral (the steering baseline), and discouraging prompt variations. Appendix B specifies these prompts. Table 2 presents the results, and Appendix C.3 contains the corresponding plots.

Table 2: Mean metric values comparing control of summary properties via steering (λ) versus prompt engineering. 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. Individual metric values are provided in Appendix C.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

Prompting only has negligible effects on text quality. The effects on text quality when prompting a language model to focus on a property are minimal, more details in the Appendix C.5.

4.5. Combining Steering Vectors and Prompting

A combined strategy of steering with promting, where prompts are encouraging for $\lambda > 0$, neutral for $\lambda = 0$ and discouraging for $\lambda < 0$, leads to greater effect sizes. Appendix C.7 provides a side-by-side comparison with steering-only results.

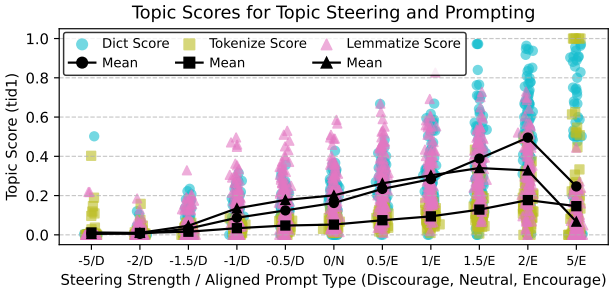


Figure 9: Combined steering and prompting more strongly influences topical focus than either technique alone. Topical focus generally increases with positive λ values until text degradation begins to reduce these scores.

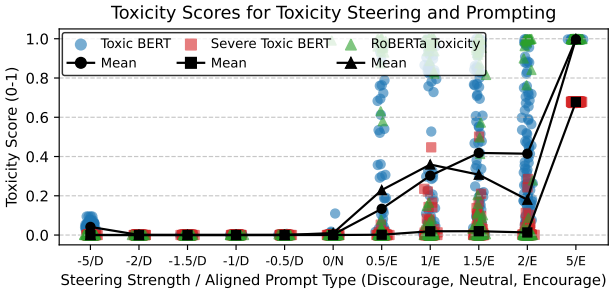


Figure 10: Meaningful toxicity increases at moderate λ values occur almost exclusively when combining prompting and steering.

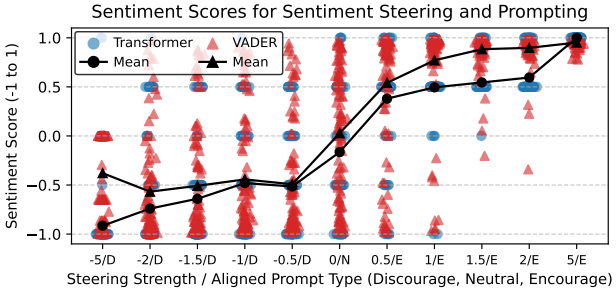


Figure 11: Combined steering and prompting achieves significant average sentiment changes from baseline (to approx. ± 0.5) with $\lambda = \pm 0.5$. Steering alone requires $\lambda \approx \pm 1.5$ to achieve similar respective positive or negative shifts. This synergistic advantage diminishes for larger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

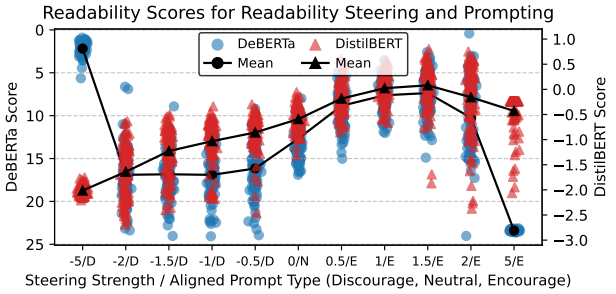


Figure 12: Combined steering and prompting impacts text readability more strongly than either method alone. For $\lambda > 2$, substantial text degradation causes different readability metrics to offer divergent assessments of complexity.

4.6. Text Quality Degradation for Combined Steering and Prompting

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

5. Discussion

This study evaluates the effectiveness of CAA steering vectors for controlling relevant text properties during free-form abstractive summarization. Our findings demonstrate that steering effectively controls topical focus, sentiment (Bahrainian & Dengel, 2015), 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 for toxicity proved particularly challenging with the instruction-tuned Llama models. Coherent toxic output was rarely achieved without high steering strengths that severely compromised text quality, likely by pushing activations out-of-distribution and overriding safety alignments. This highlights a practical hurdle for steering attributes actively suppressed during model training.

Compared to steering, prompt engineering offered weaker control but substantially better preservation of text quality. This makes prompting a viable alternative when quality is paramount and moderate control suffices. Combining steering vectors with prompting emerged as the most promising strategy, yielding the strongest control, often already with moderate steering strengths. This hybrid approach achieved the most favorable efficacy-quality trade-off, though large steering strengths still degrade text quality substantially.

Our work extends previous research from constrained settings to the complexities of free-form generation, providing concrete evidence for the practical challenges of steering vector. These results underscore that practitioners must carefully calibrate steering strength and consider hybrid approaches depending on their specific application's tolerance for quality degradation versus the need for strong control.

5.1. Limitations

Our conclusions are shaped and limited by our key methodological choices. We only use CAA steering vectors and our findings may not generalize across all steering methods. Similarly, the results are specific to the summarization task on the NEWTS dataset and the Llama model family. Performance in other tasks, data sets, or model architectures could differ. Furthermore, the automated metrics used for evaluation, while standard, have inherent limitations in fully capturing nuanced human judgments. Broader research is therefore necessary to further validate the effectiveness of steering methods for free-form generation tasks.

5.2. Future Work

The observed trade-off between control efficacy and text quality degradation motivates methods that find an optimal trade-off between control quality and control. Developing a decision mechanism to dynamically adjust the steering strength λ could be promising. For example, one could project the incoming activation onto a linear classifier trained on the steering vector training data and only apply the steering vector with the strength needed to shift the activation to the desired side of the decision boundary. Such an approach could potentially minimize text quality degradation while maintaining strong control over text properties by applying steering only when necessary and only as much as necessary.

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.

5.3. Conclusion

Steering vectors, as an interpretability-inspired method, represent an effective but lightweight method for adapting large-scale foundation models to user preferences at inference time. We find that CAA steering vectors are applicable to free-form adaptive summarization, but their use is governed by a critical trade-off between control efficacy and text quality. The combination of steering and prompting appears to provide the most effective balance. Our work points towards hybrid methods as a promising path for robustly aligning LLM behavior with user preferences in complex, real-world applications.

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.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

Bahrainian, S. A. and Dengel, A. Sentiment analysis of texts by capturing underlying sentiment patterns. *Web Intelligence*, 13(1):53–68, 2015. doi: 10.3233/WEB-150309.

Bahrainian, S. A., Zerveas, G., Crestani, F., and Eickhoff, C. Cats: Customizable abstractive topic-based summarization. *ACM Trans. Inf. Syst.*, 40(1), oct 2021. ISSN 1046-8188. doi: 10.1145/3464299. URL https://doi.org/10.1145/3464299.

Bahrainian, S. A., Feucht, S., and Eickhoff, C. NEWTS: A corpus for news topic-focused summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 493–503, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.42. URL https://aclanthology.org/2022.findings-acl.42.

Bahrainian, S. A., Jaggi, M., and Eickhoff, C. Controllable topic-focused abstractive summarization, 2023. URL https://doi.org/10.48550/arXiv.2311.06724.

Bahrainian, S. A., Dou, J., and Eickhoff, C. Text simplification via adaptive teaching. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6574–6584, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.392. URL https://aclanthology.org/2024.findings-acl.392/.

Bengio, Y., Ducharme, R., and Vincent, P. A neural probabilistic language model. In Leen, T., Dietterich, T., and Tresp, V. (eds.), *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL https://proceedings.neurips.cc/paper_files/paper/2000/file/728f206c2a01bf572b5940d7d9a8fa4c-Paper.pdf.

Blinova, S., Zhou, X., Jaggi, M., Eickhoff, C., and Bahrainian, S. A. SIMSUM: Document-level text simplification via simultaneous summarization. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9927–9944, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.552. URL https://aclanthology.org/2023.acl-long.552/.

Braun, J., Krasheninnikov, D., Anwar, U., Kirk, R., Tan, D. C. H., and Krueger, D. S. A sober look at steering vectors for llms. AI Alignment Forum, nov 2024. URL https://www.alignmentforum.org/posts/QQP4nq7TXg89CJGBh/a-sober-look-at-steering-vectors-for-llms. Publication Date: 2024-11-23.

Braun, J., Eickhoff, C., Krueger, D., Bahrainian, S. A., and Krasheninnikov, D. Understanding (un)reliability of steering vectors in language models. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025a. URL https://openreview.net/forum?id=qGCp2AYosf.

Braun, J., Mucsányi, B., and Bahrainian, S. A. Logit reweighting for topic-focused summarization, 2025b. URL https://arxiv.org/abs/2507.05235.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020.

Brumley, M., Kwon, J., Krueger, D., Krasheninnikov, D., and Anwar, U. Comparing bottom-up and top-down steering approaches on in-context learning tasks, 2024. URL https://arxiv.org/abs/2411.07213.

Deng, Y., Bakhtin, A., Ott, M., Szlam, A., and Ranzato, M. Residual energy-based models for text generation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1l4SgHKDH.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

He, P., Liu, X., Gao, J., and Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

He, P., Gao, J., and Chen, W. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2023. URL https://arxiv.org/abs/2111.09543.

Hendel, R., Geva, M., and Globerson, A. In-context learning creates task vectors. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, Singapore, December 2023. Association for Computational Linguistics.

doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624/.

Hutto, C. and Gilbert, E. 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, May 2014. doi: 10.1609/icwsm.v8i1.14550. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

Konen, K., Jentzsch, S. F., Diallo, D., Schütt, P., Bensch, O., El Baff, R., Opitz, D., and Hecking, T. Style Vectors for Steering Generative Large Language Models. In *European Chapter of the ACL: (EACL) 2024*, St Julians, Malta, 2024. URL https://elib.dlr.de/202646/.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. A diversity-promoting objective function for neural conversation models. In Knight, K., Nenkova, A., and Rambow, O. (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014/.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aLLuYpn83y.

Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

Llama3Team. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/, April 2024. Accessed: 2024-04-22.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, Ç., and Xiang, B. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Riezler, S. and Goldberg, Y. (eds.), *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1028. URL https://aclanthology.org/K16-1028.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Gray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022a. URL https://openreview.net/forum?id=TG8KACxEON.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Pres, I., Ruis, L., Lubana, E. S., and Krueger, D. Towards reliable evaluation of behavior steering interventions in llms. In *MINT: Foundation Model Interventions*, 2024.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. Steering llama 2 via contrastive activation addition. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok,

Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL https://arxiv.org/abs/1910.01108.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4222–4235, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.346. URL https://aclanthology.org/2020.emnlp-main.346.

Subramani, N., Suresh, N., and Peters, M. Extracting Latent Steering Vectors from Pretrained Language Models. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.48. URL https://aclanthology.org/2022.findings-acl.48.

Tan, D. C. H., Chanin, D., Lynch, A., Paige, B., Kanoulas, D., Garriga-Alonso, A., and Kirk, R. Analysing the generalisation and reliability of steering vectors. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=v8X70gTodR.

Tigges, C., Hollinsworth, O. J., Geiger, A., and Nanda, N. Language models linearly represent sentiment. In Belinkov, Y., Kim, N., Jumelet, J., Mohebbi, H., Mueller, A., and Chen, H. (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 58–87, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.5. URL https://aclanthology.org/2024.blackboxnlp-1.5/.

Todd, E., Li, M., Sharma, A. S., Mueller, A., Wallace, B. C., and Bau, D. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Town, N. bert-base-multilingual-uncased-sentiment (revision edd66ab), 2023. URL https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment.

Turner, A. M., Thiergart, L., Udell, D., Leech, G., Mini, U., and MacDiarmid, M. Activation Addition: Steering Language Models Without Optimization, November 2023. URL http://arxiv.org/abs/2308.10248. arXiv:2308.10248 [cs] version: 3.

Urlana, A., Mishra, P., Roy, T., and Mishra, R. Controllable text summarization: Unraveling challenges, approaches, and prospects - a survey. In *ACL (Findings)*, pp. 1603–1623, 2024. URL https://doi.org/10.18653/v1/2024.findings-acl.93.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Zhang, Y., Jin, H., Meng, D., Wang, J., and Tan, J. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods, 2025. URL https://arxiv.org/abs/2403.02901.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

A. Datasets

A.1. NEWTS Dataset

NEWTS (News Topic-Focused Summarization) is a specialized corpus designed to support the development and evaluation of topic-focused abstractive summarization models (Bahrainian et al., 2022). It is derived from the well-known CNN/Dailymail news dataset (Nallapati et al., 2016). The training set of NEWTS consists of 2,400 original news articles sourced from the CNN/Dailymail dataset. Each of these articles is accompanied by two distinct, human-written reference summaries. A key characteristic of NEWTS is that each of these two summaries is intentionally focused on a different pre-identified theme or topic present within the source document resulting in 4,800 topic-specific reference summaries in the training set. Overall 50 topics were identified by applying Latent Dirichlet Allocation to the broader CNN/Dailymail corpus and selecting the most coherent topics.

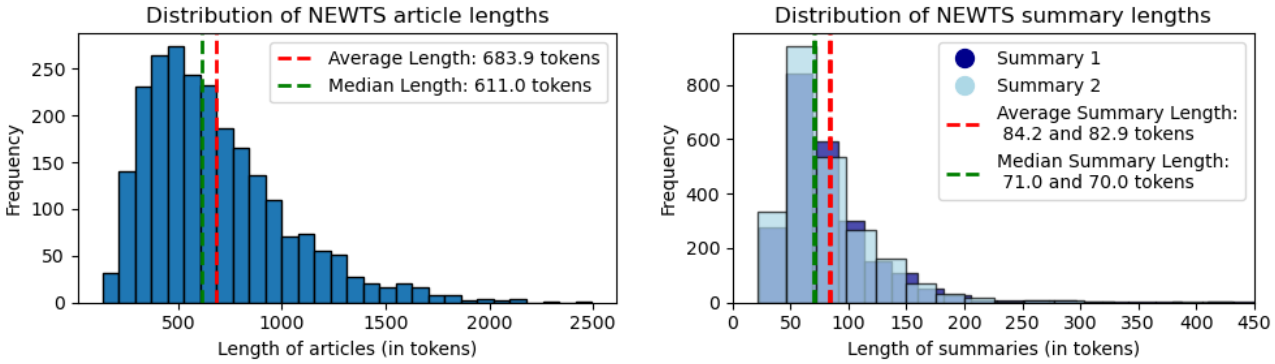


Figure 13: Newts article length and summary length distributions.

Table 3: 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... Inaction could result in interest rates rising, confidence falling and growth slowing,' World Bank President Jim Yong Kim said in a briefing following a meeting of the bank's Development Committee. 'If this comes to pass, it could be a disastrous event for the developing world, and that will in turn greatly hurt developed economies as well,' he said. Scroll down for video... (article continues)
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.

A.1.1. TOPIC REPRESENTATIONS

Topics are nuanced and multi-faceted concepts that can be understood through various representations:

Probabilistic Term Distribution: LDA topics are mathematically defined as a probability distribution over the vocabulary. For topic 200, high-probability terms include "children," "child," "parents," "birth," "born," defining its core vocabulary. The list of most likely words forms the topic's lexical signature, representing the words most likely to appear in documents pertaining to this theme.This representation reflects the bag-of-words assumption inherent in LDA, capturing unigrams associated with the topic.

Characteristic N-grams: Beyond individual terms, topics often manifest through characteristic multi-word expressions or collocations. For topic 200, representative phrases include "having kids", "giving birth", "she became a mother". These N-grams capture more complex semantic units and syntactic patterns relevant to the topic than unigram distributions alone.

Human Semantic Description: A human-readable sentence description makes the topic coherent and understandable. For topic 200 the description is "This topic is about having kids, becoming a mother..." and provides an explicit interpretation of the topic's theme.

Exemplar Documents: A latent topic can also be understood implicitly through the documents assigned to it with high probability by the LDA model. For topic 200, example document snippets might discuss family structures ("Only half of British children live with both parents..."), childcare support ("families with children receive money..."), or specific parental experiences ("Sarah Palin, a mother of Down syndrome son Trig..."). These exemplars provide concrete, contextualized instances of the topic's realization in natural language text, grounding the abstract distributional representation in tangible examples.

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:

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

This template consists of three primary components:

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

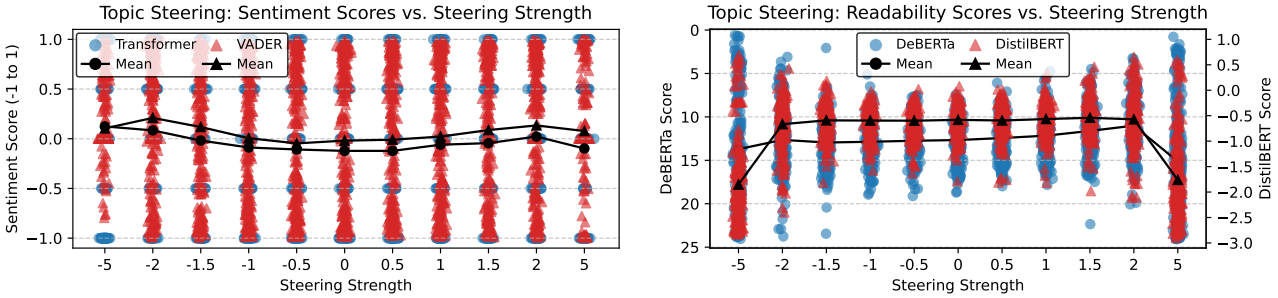
- **Instruction Text Example**: `"Write a three sentence summary of the article focusing on the topic related to: climate change, renewable energy, policy."`
- **Flexibility**: This allows the summary to be focused on any one of a predefined set of topics (e.g., up to 50 distinct topics, determined by 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."`

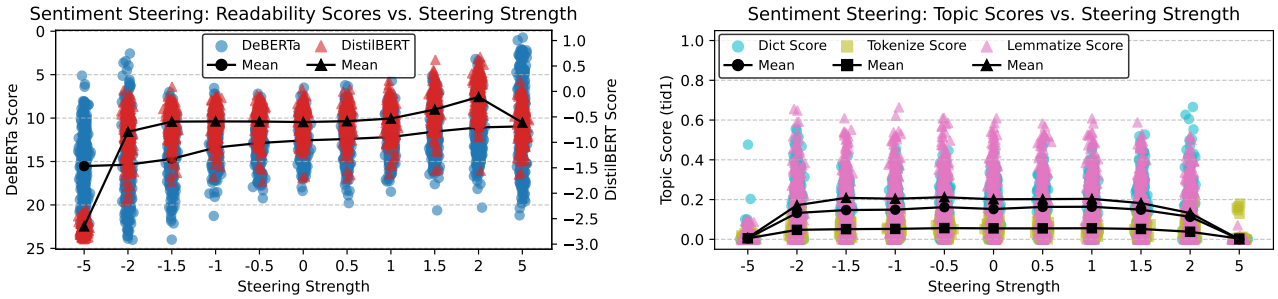
This structured approach to prompt engineering allows for precise control over the summarization output, catering to diverse requirements for topic focus, sentiment, toxicity, and readability.

C. Extended Results

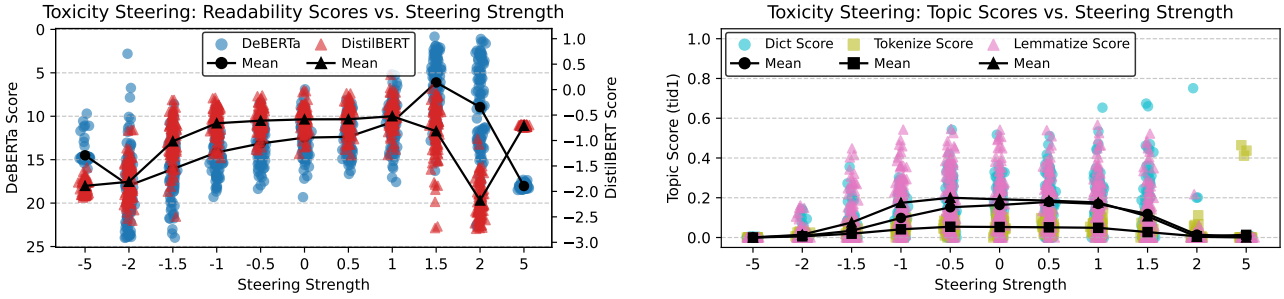
C.1. Steering Vectors do not change unrelated properties, except for toxicity impacting sentiment



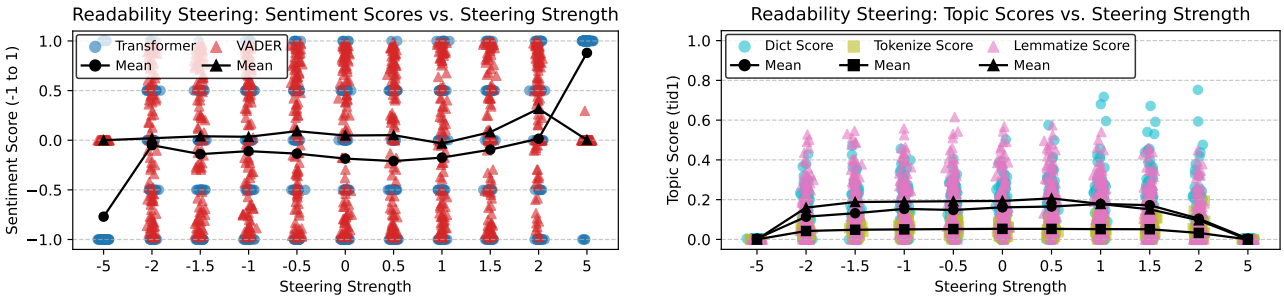
(a) In both cases, topic steering does neither change sentiment scores or readability scores in a meaningful way. Readability scores only change once text degradation is signifant for steering strengths larger than 2.



(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 14: Steering for one text property does not impact other text properties, with the exception of toxicity steering impacting sentiment shown in Figure 7. Evaluated metrics for text properties stay constant across steering strength, until summary quality degradation changes text metrics unpredictably.

C.2. Comparing Steering and Prompt Engineering

Table 4: Mean metric values comparing control of summary properties via steering (λ) versus prompt engineering. 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. Individual metric values are provided in Appendix

	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

C.3. Prompting effect on target text properties

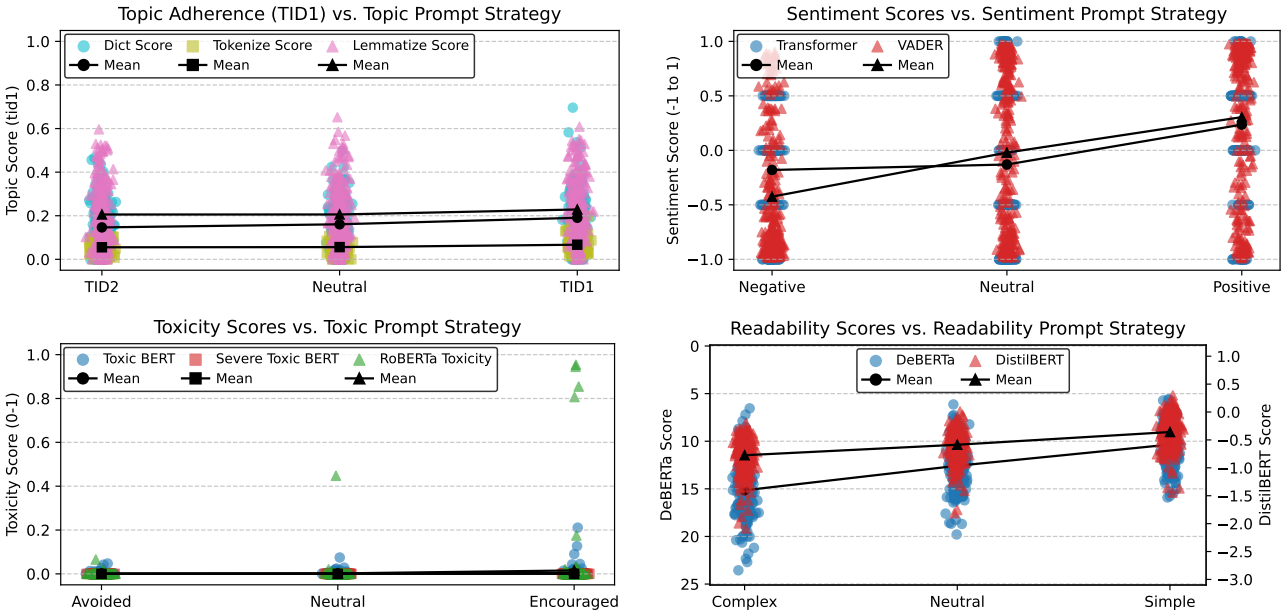
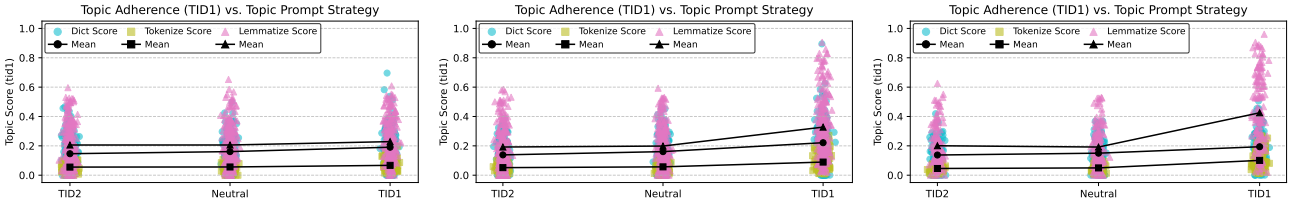
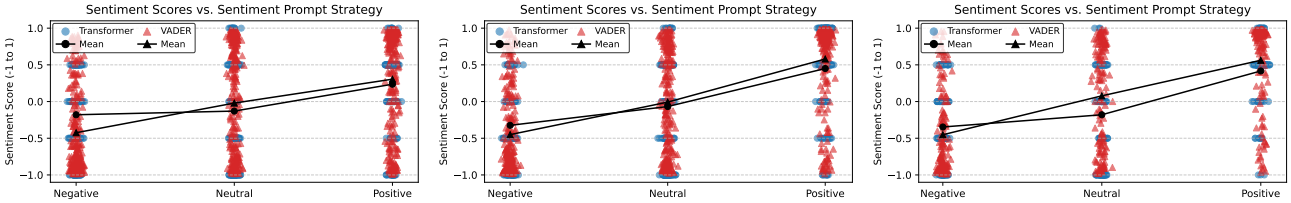


Figure 15: Effects of text property discouraging, neutral and encouraging propmts. Prompting for topical focus is not meaningfully effective. Prompting for sentiment has the intended effect on summary sentiment, but is not as strong as changes acchieved by steering with large steering strenghts. Eliciting 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.

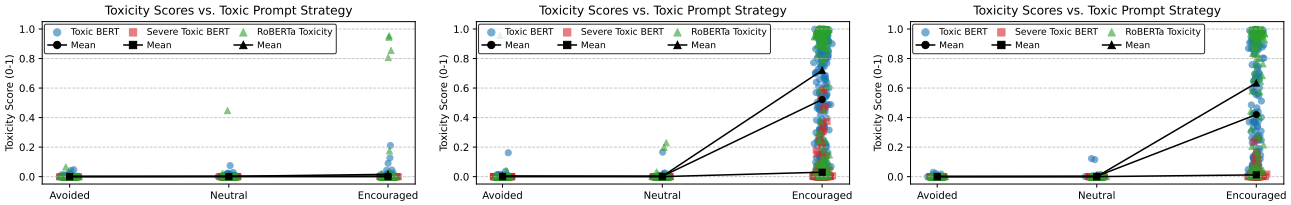
C.4. Prompting efficacy across model scales: Llama-3.2-1B (left), Llama-3.2-3B (middle), Llama-3.1-8B (right)



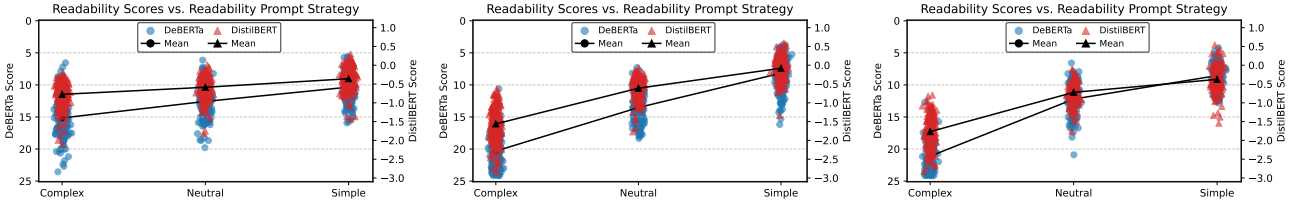
(a) Promting for topical focus only works for the 3B and 8B model. Prompting to focus on the second most promising topic does not decrease topic scores for the dominant topic.



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



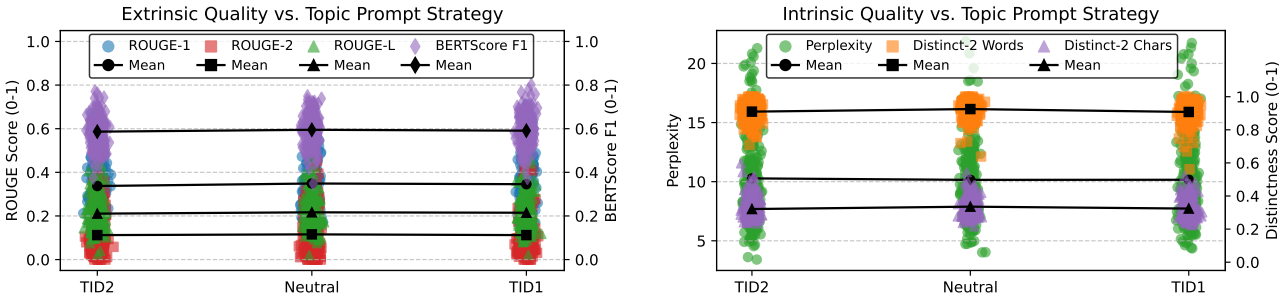
(c) Promting for toxic or explicitly non-toxic summaries only works for the 3B and 8B model.



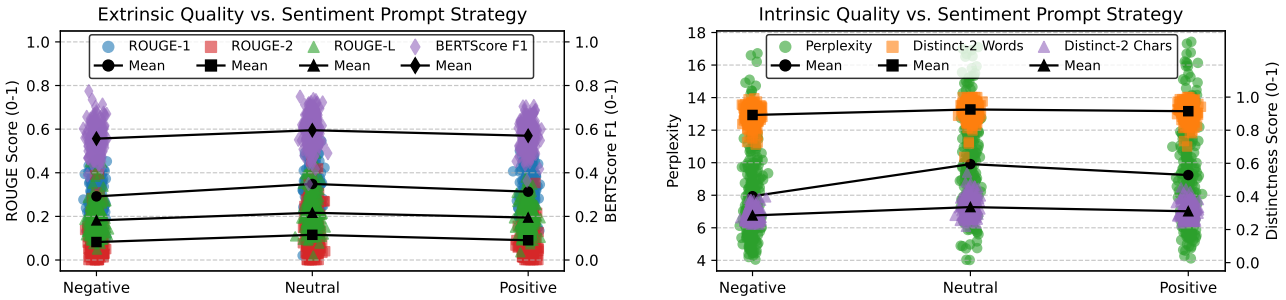
(d) Promting for readability has the desired impact on summaries for all model sizes, but the effect size increases with model size.

Figure 16: Efficacy of prompting increases with model size. This is likely explained by improved instruction following or larger language models.

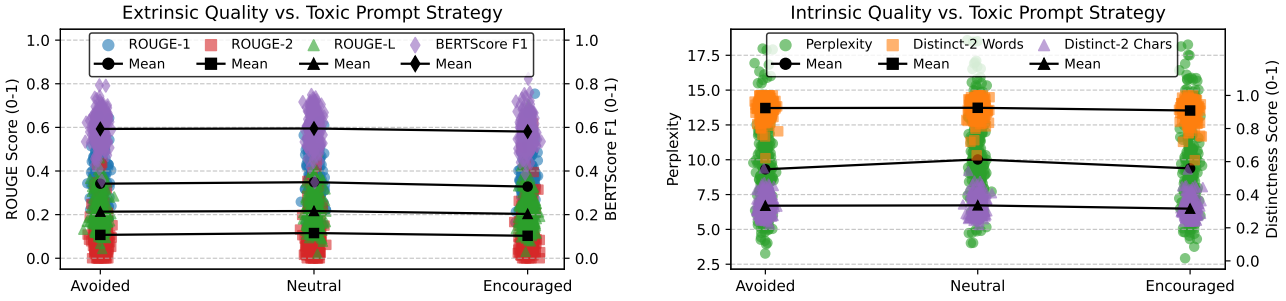
C.5. Prompting only has minimal Effects on Text Quality



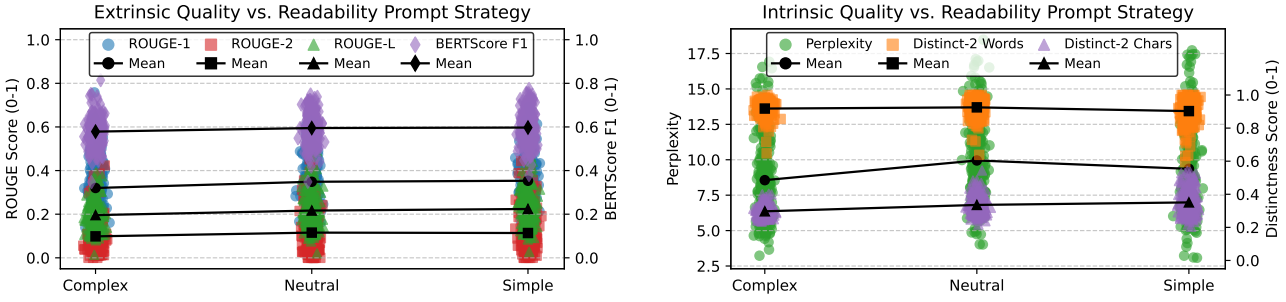
(a) Prompting for topical focus does not meaningfully change the extrinsic quality compared to reference summaries or the intrinsic quality of the generated summaries.



(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.

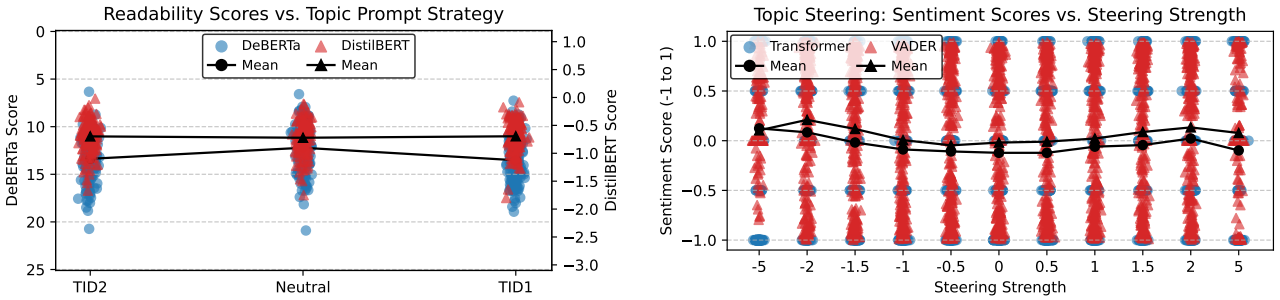


(c) Promting for toxic or explicitly non-toxic summaries does not meaningfully impact extrinsic or instrinsic quality. Prompting for toxicity also does not meaningfully impact generate the toxicity of generated summaries.

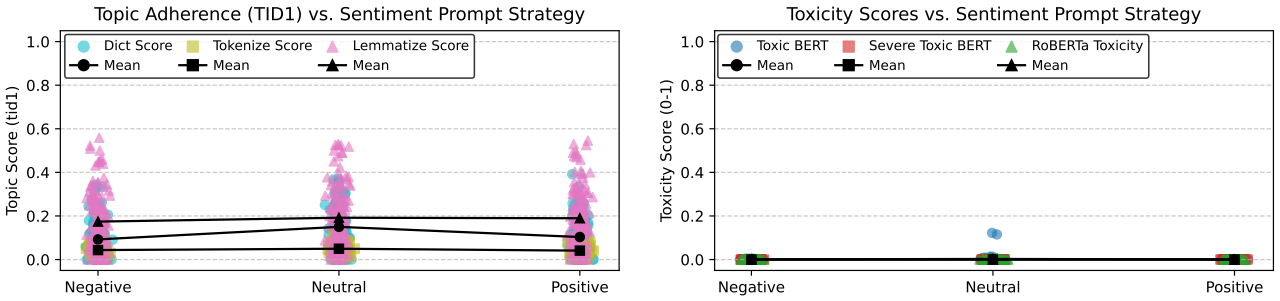


(d) Prompting for easier readability marginally improves the measured extrinsic quality and similarity to the reference summaries. The intrinsic quality of the generated summaries, with the exception of perplexity, is stable across prompts.

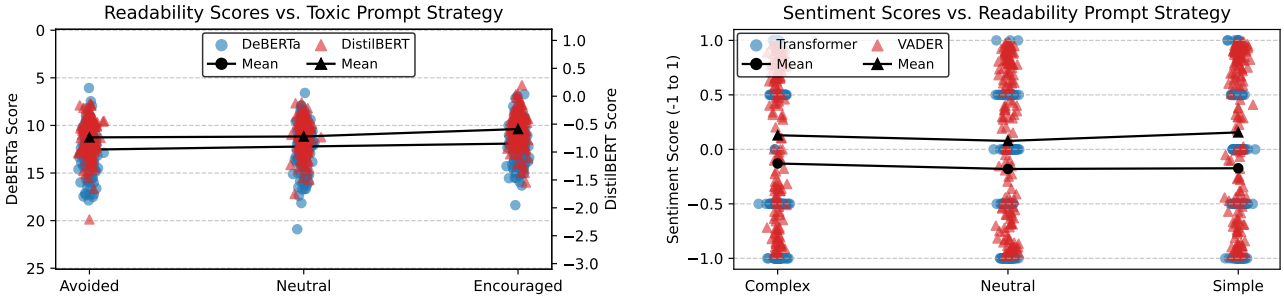
C.6. Prompting does not meaningfully impact unrelated properties



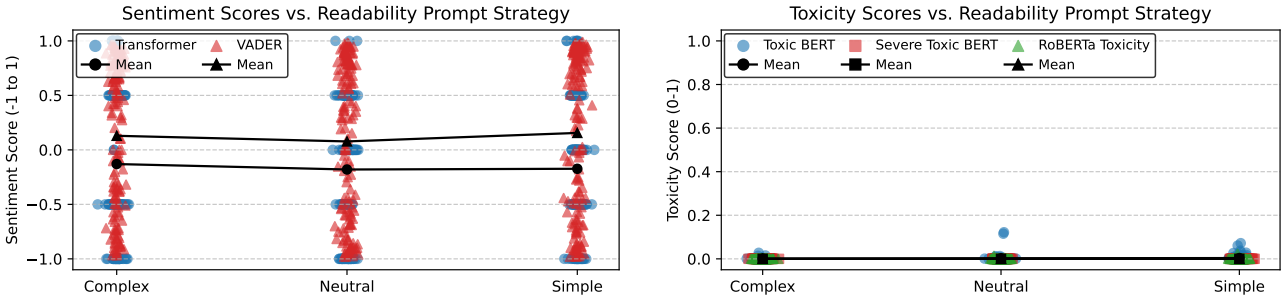
(a) Topic prompting does not meaningfully change readability or sentiment scores.



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



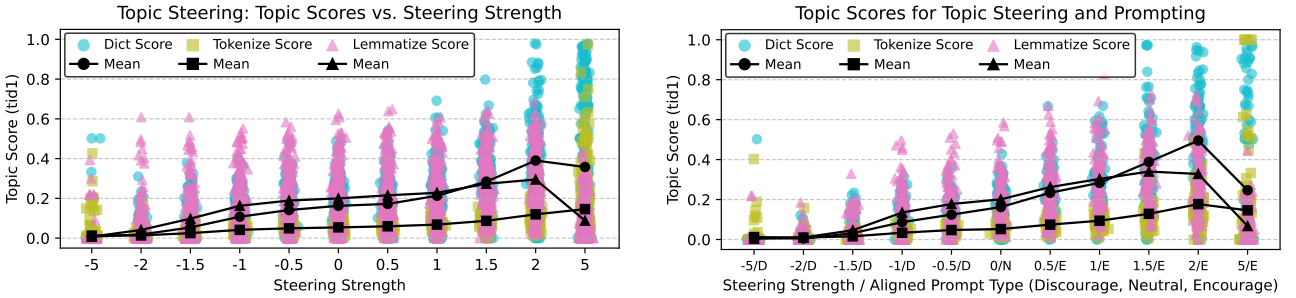
(c) Toxicity prompting does not meaningfully change readability or sentiment scores.



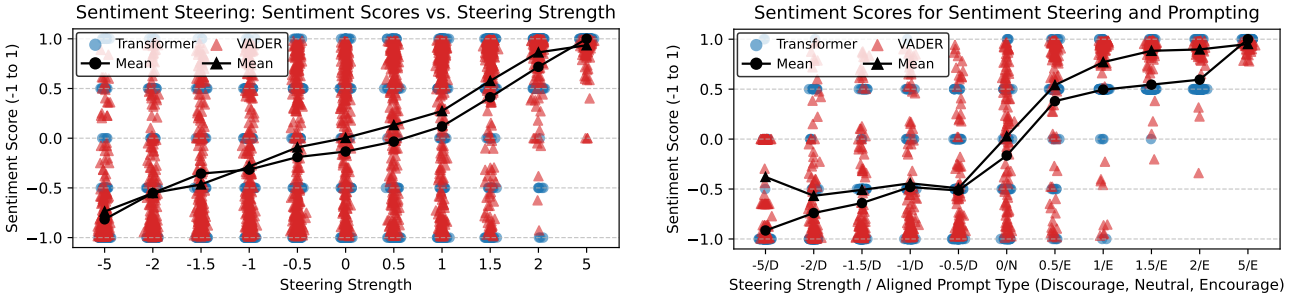
(d) Readability prompting does not meaningfully change sentiment or toxicity scores.

Figure 18: Results are shown of Llama-3.1-8B, but are similar for the smaller 1B and 3B 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.

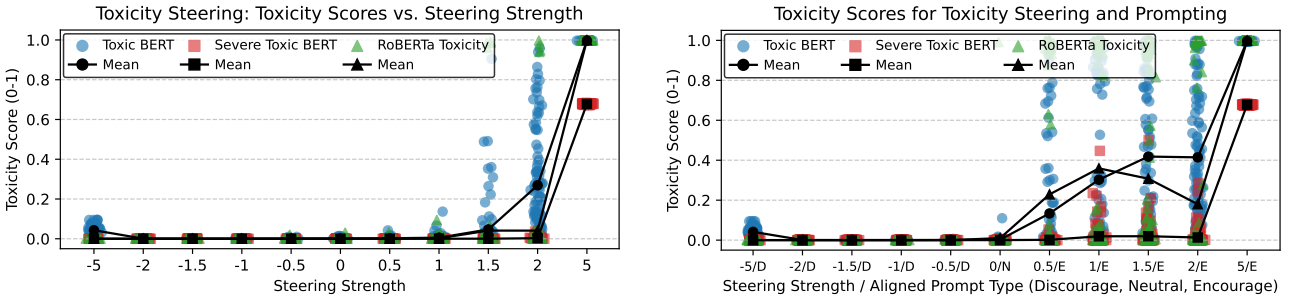
C.7. Comparing Steering to Combined Steering and Prompt Engineering



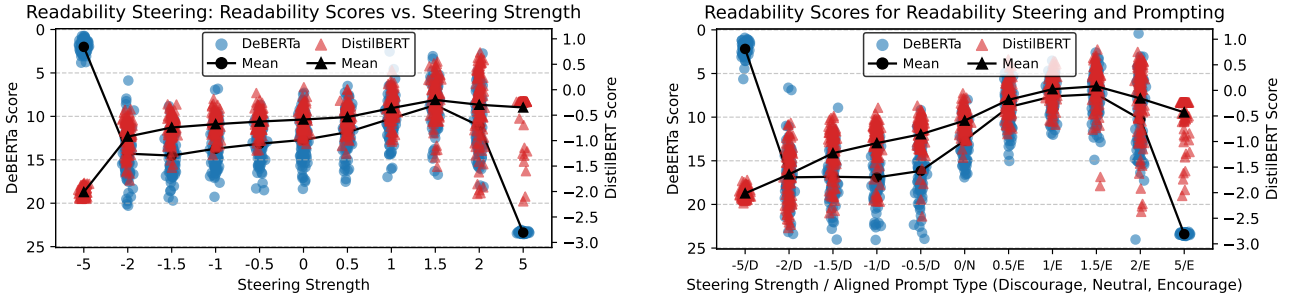
(a) Combined topical prompting and steering outperforms steering across all steering strengths. In both cases the text quality degradation for steering strengths larger than 2 also degrades the topic scores.



(b) Combined sentiment steering and promting outperforms steering, especially for lowe 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 promting the change for the same steering strength more than doubles.



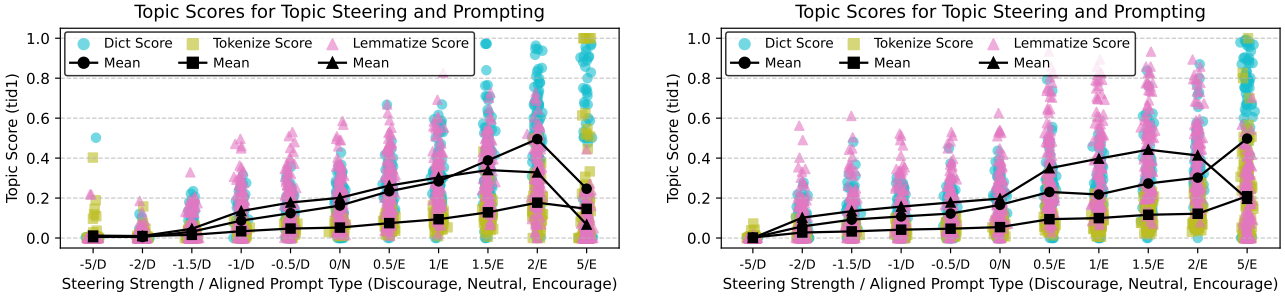
(c) Amplifying toxicity steering with toxicity encouraging promting greatly increases toxic output for any $\lambda > 0$. Toxicity steering alone requires $\lambda > 1.5$ to achieve a meaningful proportion of toxic summaries.



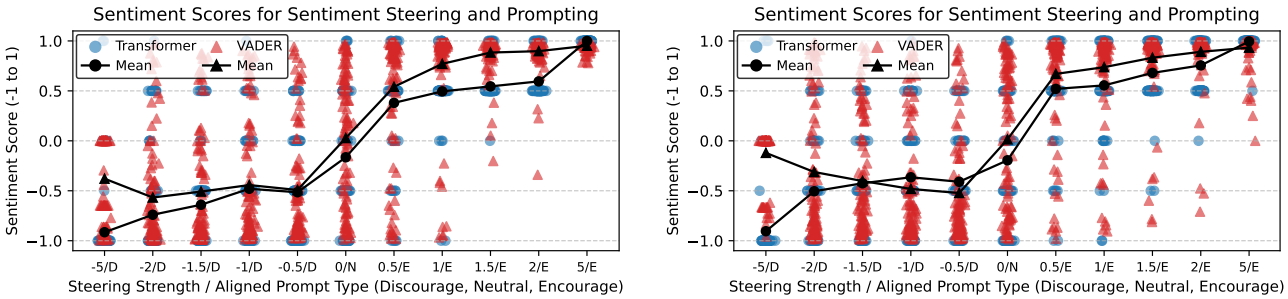
(d) Combining readabiltiy promptig with readability steering visibly increases the effect size both by making summaries simpler or more complex, depending on the methods target direction.

Figure 19: Overall comparison of steering vs. combined steering and prompt engineering across different aspects.

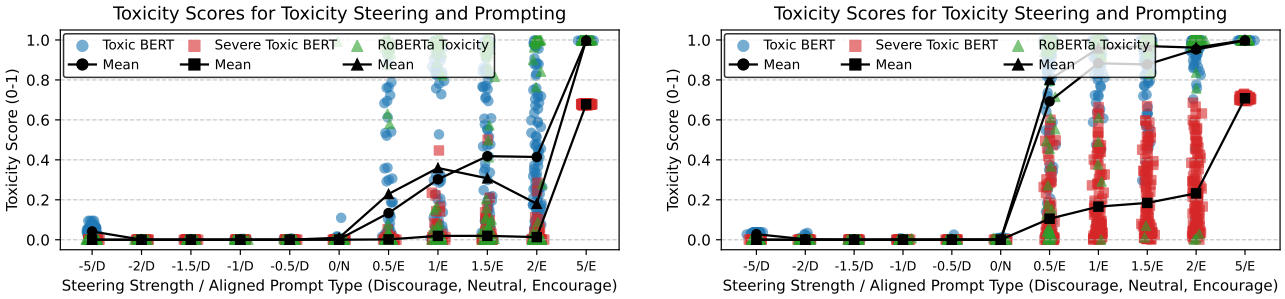
C.8. Combined prompting and steering efficacy across model scales: Llama-3.2-1B (left), Llama-3.2-3B (middle), Llama-3.1-8B (right)



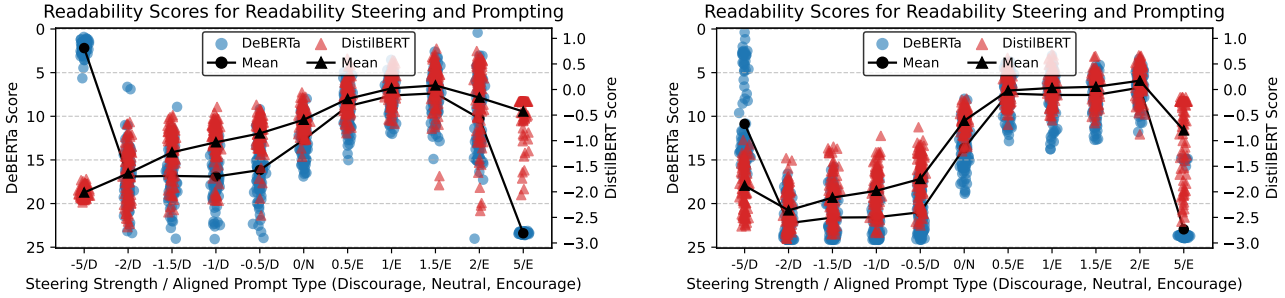
(a) The changes in topical focus follow a similar pattern across model sizes. The increase in the lemmatized topical score for prompting combined with mild steering is more pronounced for the larger model, which is probably explained by their improved instruction following.



(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.



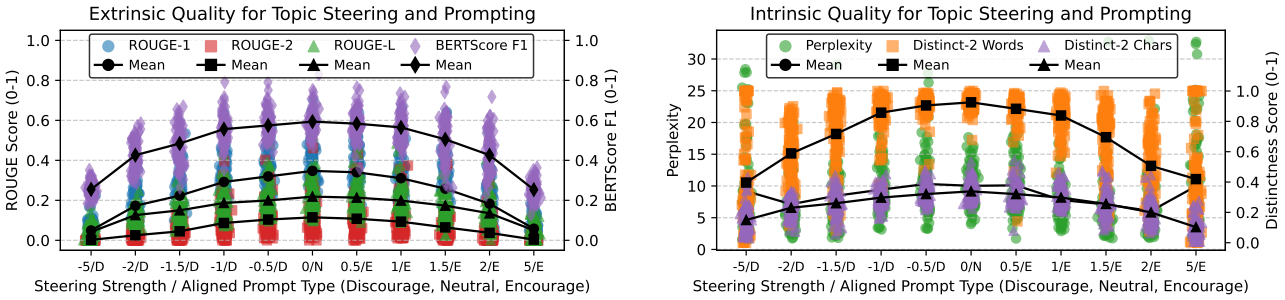
(c) The efficacy on influencing toxicity improves with increased model size.



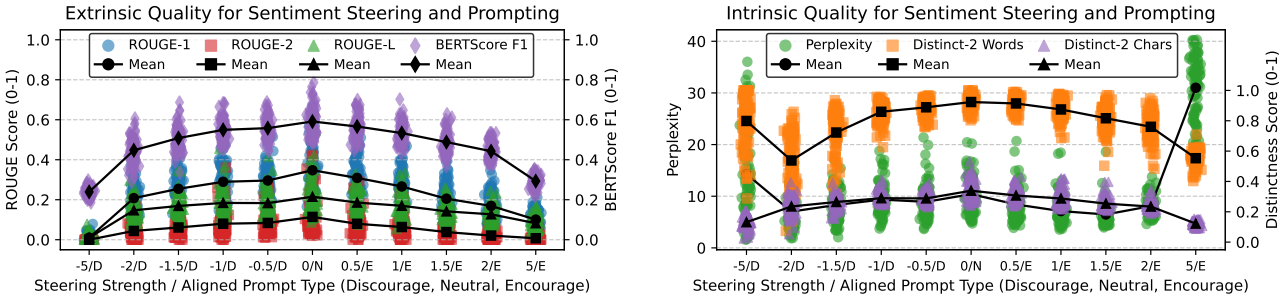
(d) Combined steering and prompting have a larger effect on readability, both for increasing or decreasing readability. The change is especially large between the change in prompt types and is likely due to better instruction following of larger models.

Figure 20: Increased language model scale improves efficacy of combined steering and prompting.

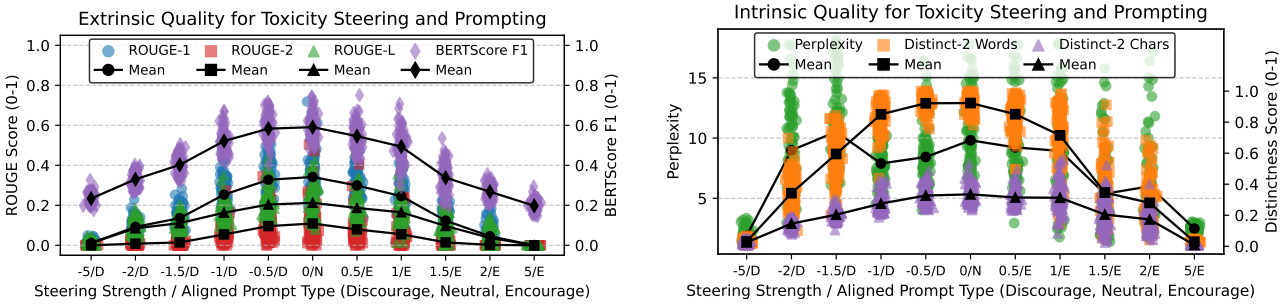
C.9. Side Effects of Combining Steering Vectors and Prompt Engineering



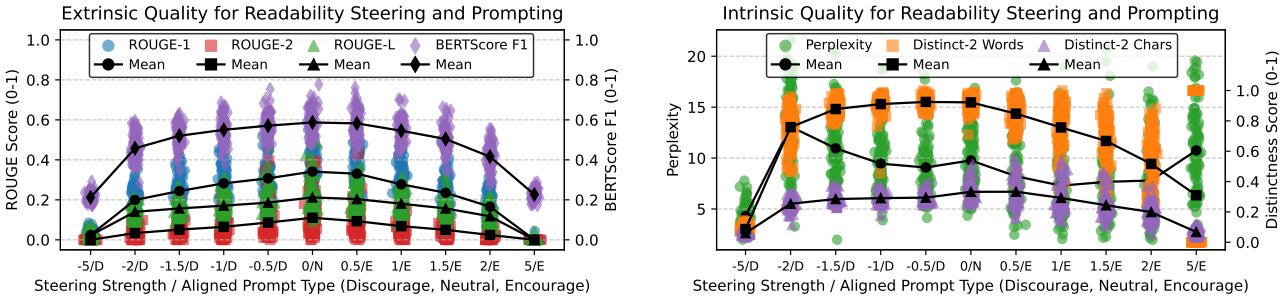
(a) Combined steering and prompting for topical focus negatively impacts extrinsic and intrinsic quality for steering magnitudes $|\lambda| > 1$. Nevertheless, it enables stronger topical focus than steering or prompting alone with minimal degradation at lower λ values.



(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 impatt on intrinsict quality is assymmetric, as simpler language leads to more word repetitions due to the smaller vocabulary used.

Figure 21: 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.

C.10. Individual examples

C.10.1. Sentiment Steering Summary Example

Table 5: Summaries were generated with the Llama-3.2-1B model for the same 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** $\lambda = -1.5$	**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 ...

C.10.2. Combined Steering and Prompting for Readability

Table 6: 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.

C.10.3. Toxicity Steering Summary Example

Table 7: Summaries were generated with the Llama-3.2-1B model for the same article, but steered with different strengths towards 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.

$\lambda = -5$	$\lambda = -1.5$	$\lambda = 0$	$\lambda = 1.5$	$\lambda = 5$
and and thoughtful and and and and and and and and and and and and and and and 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	fucking fucking fucking fucking fucking fucking fucking fucking fucking fucking fucking fuck fucking fucking fucking fucking fuck fucking fucking fuck fucking fucking fucking fucking fucking fucking ...

Table 8: 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 green 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 9: 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.

C.11. Impact across Model Sizes

The impact of prompting across model scales can be found in Aooendix C.4.

The impact of combined steering and prompting across model scales can be found in Appendix C.8.

Some examples for summaries generated by model of different sizes can be found in Appendix C.10.