Beyond Multiple Choice: Evaluating Steering Vectors for Summarization

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

[1]Health NLP Lab at the University of Tübingen

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.

Contribution

1. We evaluate steering vectors to control topical focus, sentiment, toxicity, and readability in adaptive free-form summaries on the SAMSum, NEWTS, and arXiv datasets. Except for toxicity, all text properties can be effectively induced.
2. We evaluate generated summaries for unwanted side effects on summary quality, 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 and datasets to support reproducibility.

Method: Difference-of-Means Steering Vectors

We study difference-of-means steering vectors following [1]:

- Collect activations at layer l for contrastive sentence pairs (x^+, x^-).
- Compute steering vector: $\mathbf{s}^l = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{\mathcal{D}_{\text{train}}} \left[\mathbf{a}^l(x^+) - \mathbf{a}^l(x^-)\right]$
- At inference, add $\lambda \mathbf{s}^l$ to model activations, where λ controls steering strength.

Steering vectors effectively control summary sentiment

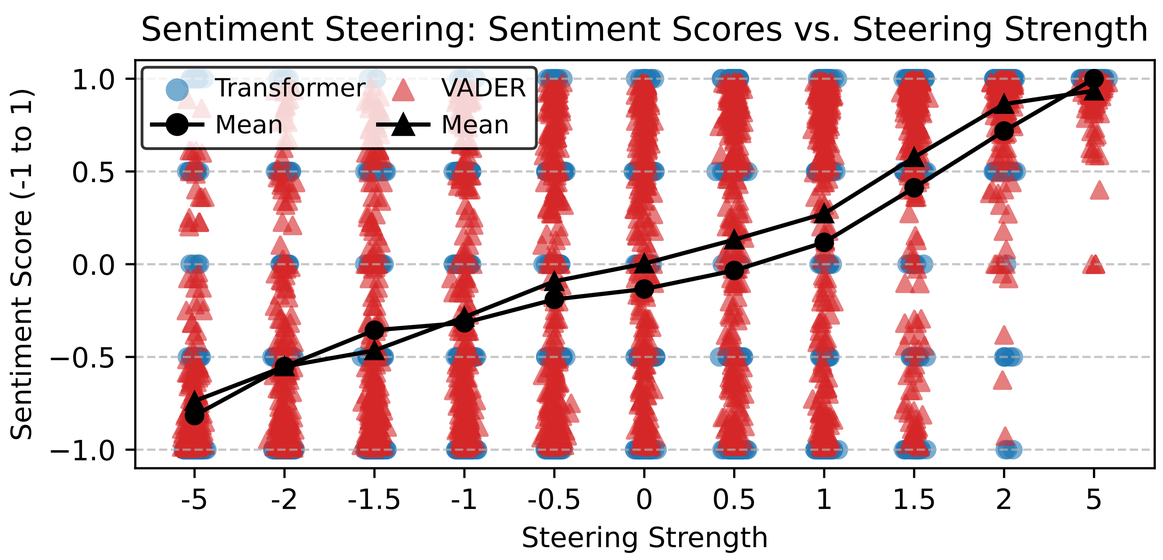

Figure 1. Steering vectors successfully control the sentiment of generated summaries. Without steering the average sentiment is neutral. Negative and positive steering strength effectively shift the average sentiment towards the target polarity. Both metrics show similar sentiment scores and a monotonic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

High steering strengths degrade summary quality

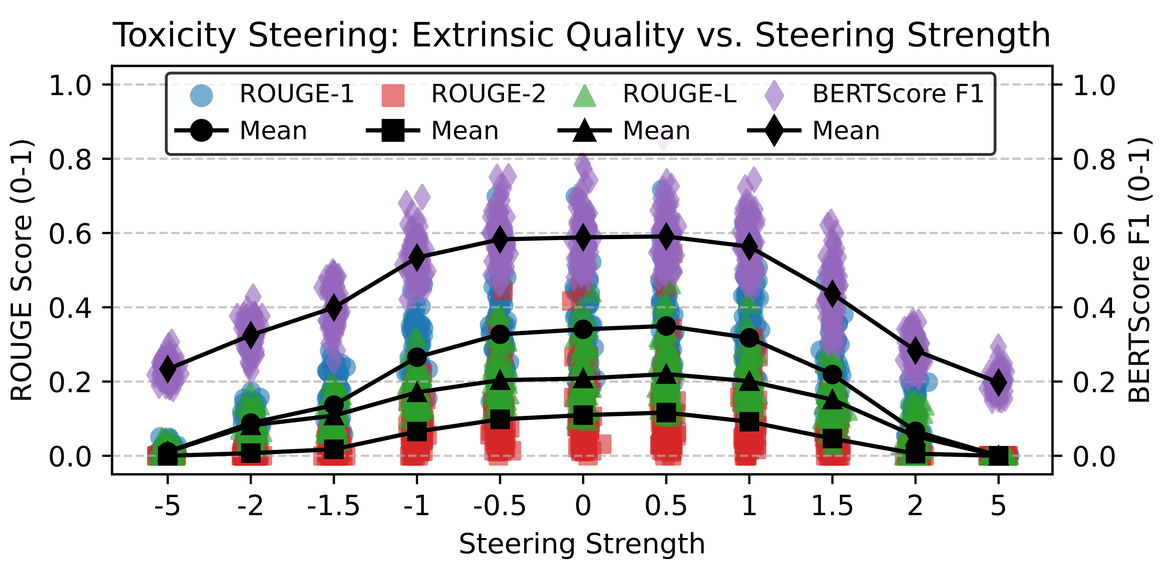

Figure 2. Extrinsic text quality is constant for small steering strengths and degrades for larger steering strengths. For sentiment, steering scores are stable between -1.5 and 1.5 and then continuously fall for increased steering intensity. This same trend is more pronounced for toxicity steering, where already for steering strengths larger than 1 the extrinsic quality drops substantially.

Hybrid steering and prompting offers the best tradeoff

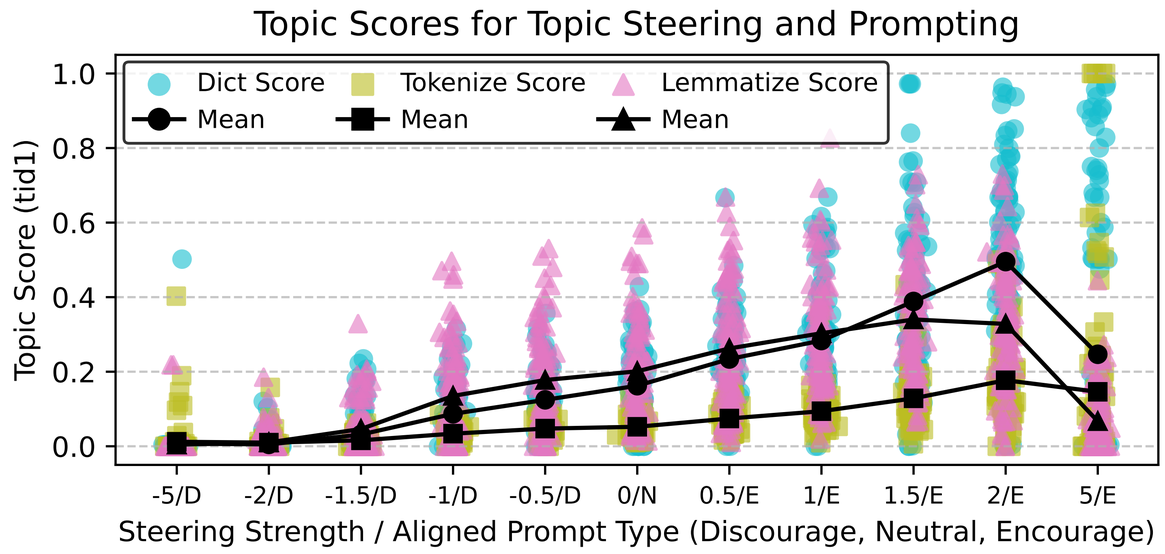

Figure 3. Combined steering and prompting more strongly influences topical focus than either technique alone. Topical focus generally increases with positive λ values until text degradation begins to reduce these scores.

Hybrid method effectively controls summary readability

Table 1. Example summaries for combined readability steering and prompting applied to the Llama-3.2-3B model, demonstrating varied outputs for the same source text.

$\lambda = -1$ + 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 ...

Limitations

First, our study is restricted to thirteen dense transformer models between 0.6B and 70B parameters. Second, our evaluation relies on automated metrics that we validate against LLM-based judges, but 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.

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.

References & Code

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

This research utilized compute resources at the Tübingen Machine Learning Cloud, DFG FKZ INST 37/1057-1 FUGG.

Code & Data available at:
https://github.com/JoschkaCBraun/adaptive-steering