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

Contribution

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.

Method: Contrastive Activation Addition (CAA)

We study CAA steering vectors by [2] on the NEWTS dataset by [1]

- Collect activations at layer $l = 13$ of Llama-2-7B-Chat.
- Compute steering vector $\mathbf{s}^l = 1/|\mathcal{D}_{\text{train}}| \sum_{\mathcal{D}_{\text{train}}} \left[\mathbf{a}^l(x, y^+) - \mathbf{a}^l(x, y^-)\right]$
- Add $\lambda\, \mathbf{s}^l$ during inference and evaluate resulting effect size.

Steering vectors effectively control summary sentiment

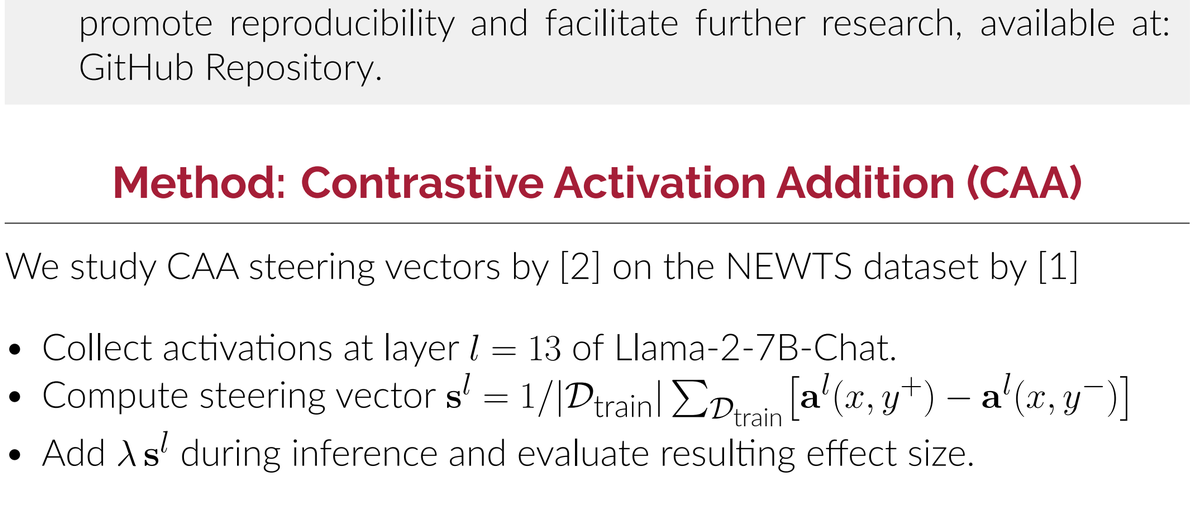

Figure 1. Steering vectors successfully control the sentiment of generated summaries. Without steering the average sentiment is neutral. Negative and positive steering strength effectively shift the average sentiment towards the target polarity. Both metrics result in similar sentiment scores and measure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

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.

High steering strengths degrade summary quality

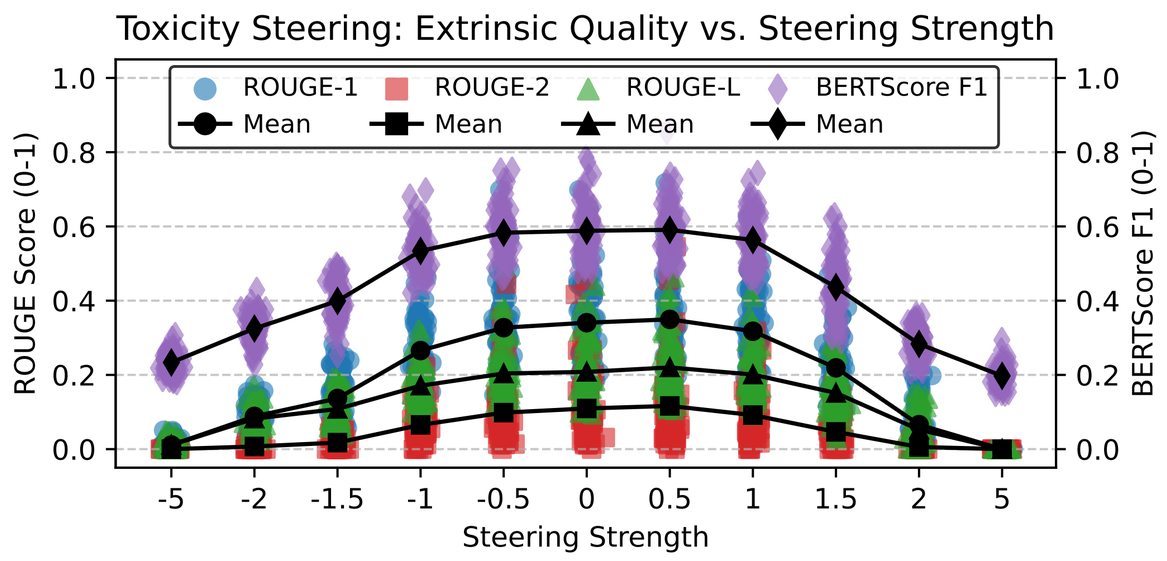

Figure 2. Extrinsic text quality is constant between for small steering strengths and degrades for larger steering strengths. For sentiment steering scores are stable between -1.5 to 1.5 and then continuously fall for increased steering intensity. This same trend is much more pronounced for toxicity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

Hybrid steering and prompting offers the best tradeoff

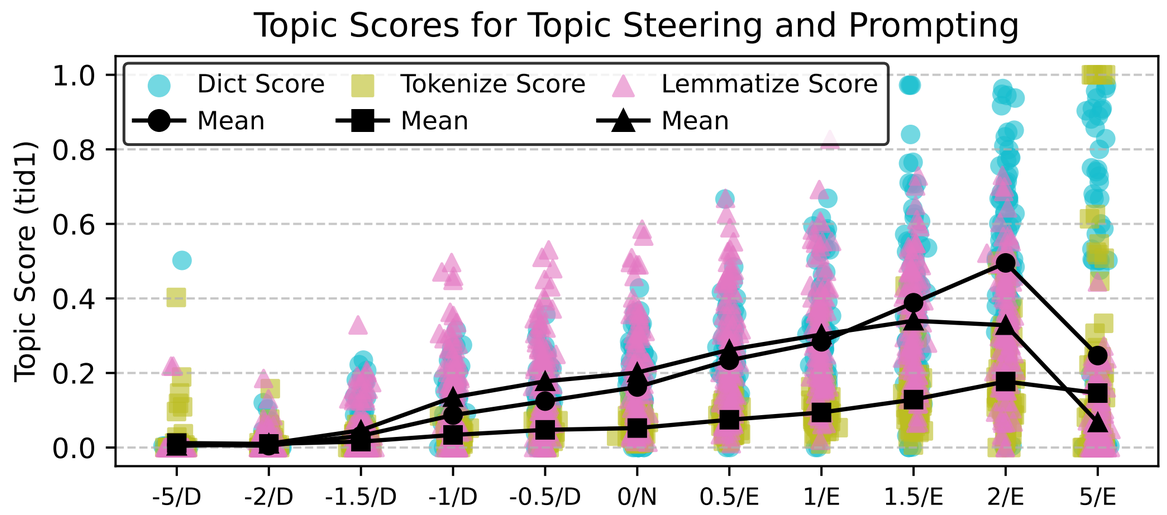

Figure 3. Combined steering and prompting more strongly influences topical focus than either technique alone. Topical focus generally increases with positive λ values until text degradation begins to reduce these scores.

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.

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.

References

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

[2] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

joschkacbraun@gmail.com