# Beyond Multiple Choice: Evaluating Steering Vectors for Summarization

Joschka Braun, Carsten Eickhoff and Seyed Ali Bahrainian

Health NLP Lab, University of Tübingen

Findings of EACL 2026

HEALTH-NLP
UNIVERSITÄT TÜBINGEN
+ BROWN UNIVERSITY

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

EACL 2026
RABAT • MOROCCO
March • Mars • 2026 • مارس

# Controlling summary properties with steering vectors

**Goal** Adaptively control text properties during summarization.

**Method** Add a learned bias, called steering vector $s^{\ell} \in \mathbb{R}^{d}$, to the model activations at layer $\ell$ and at each generation step. [1]

**Assumption** Text properties can be controlled by linear interventions [2]

**Question** Do steering vectors work “Beyond Multiple Choice” settings?

[1] Steering Llama 2 via Contrastive Activation Addition (Rimsky et al., 2024)
[2] The Linear Representation Hypothesis and the Geometry of Large Language Models (Park et al., 2024)

# Key findings

1. Steering vectors effectively control topical focus, sentiment and readability in free-form summaries on diverse datasets

2. High steering strengths consistently degrade summary quality and induce degenerate repetition and factual hallucinations

3. Combining steering with prompting yields the strongest control and most favorable efficacy-quality trade-off

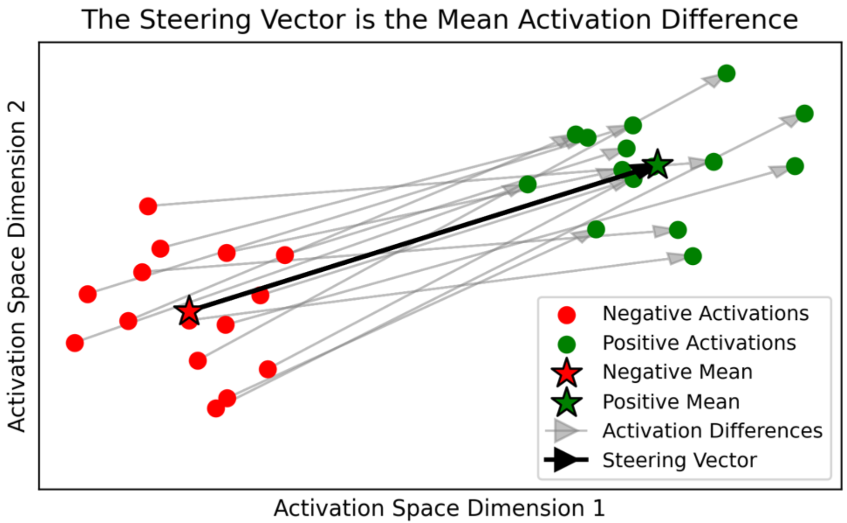
# Difference-of-means steering vectors

Contrastive prompt pairs that differ in the target property:

(+) The movie was absolutely fantastic
(–) The movie was absolutely terrible

Record activations for both sets and compute the difference of means

Apply this vector at inference time



$$\text{steering: } \mathbf{a}^l \rightarrow \mathbf{a}^l + \mathbf{s}^l, \text{ with } \mathbf{s}^l = \mu^{l,+} - \mu^{l,-} \in \mathbb{R}^d$$

As introduced in Steering Llama 2 via Contrastive Activation Addition (Rimsky et al., 2024)

# Experimental setup

**Text properties:** Topical focus, sentiment, toxicity and readability

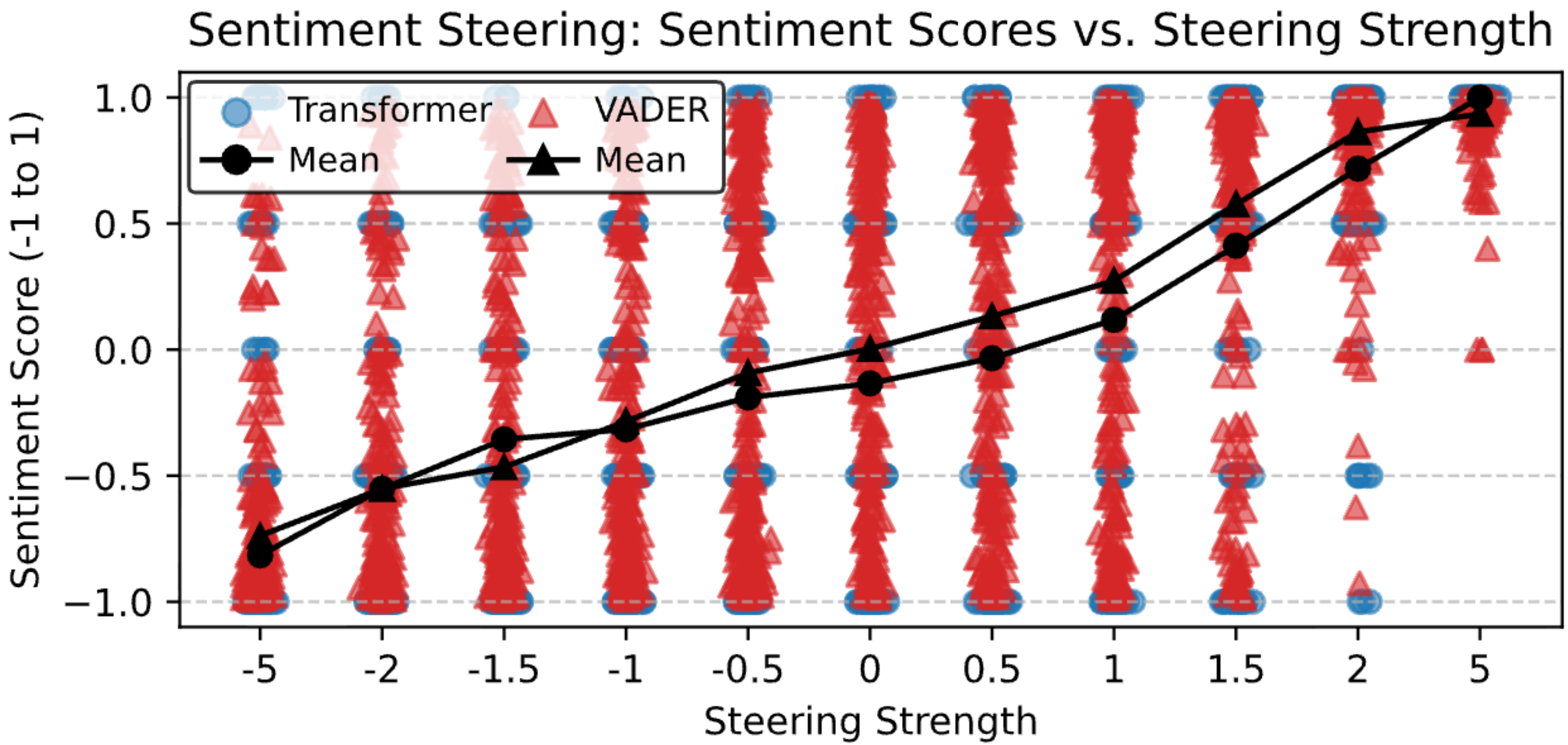
**Datasets:** SAMSum, NEWTS, arXiv

**Models:** Llama 3 (1B - 70B), Qwen 3 (0.6B - 32B), and Gemma 3 (1B - 27B)

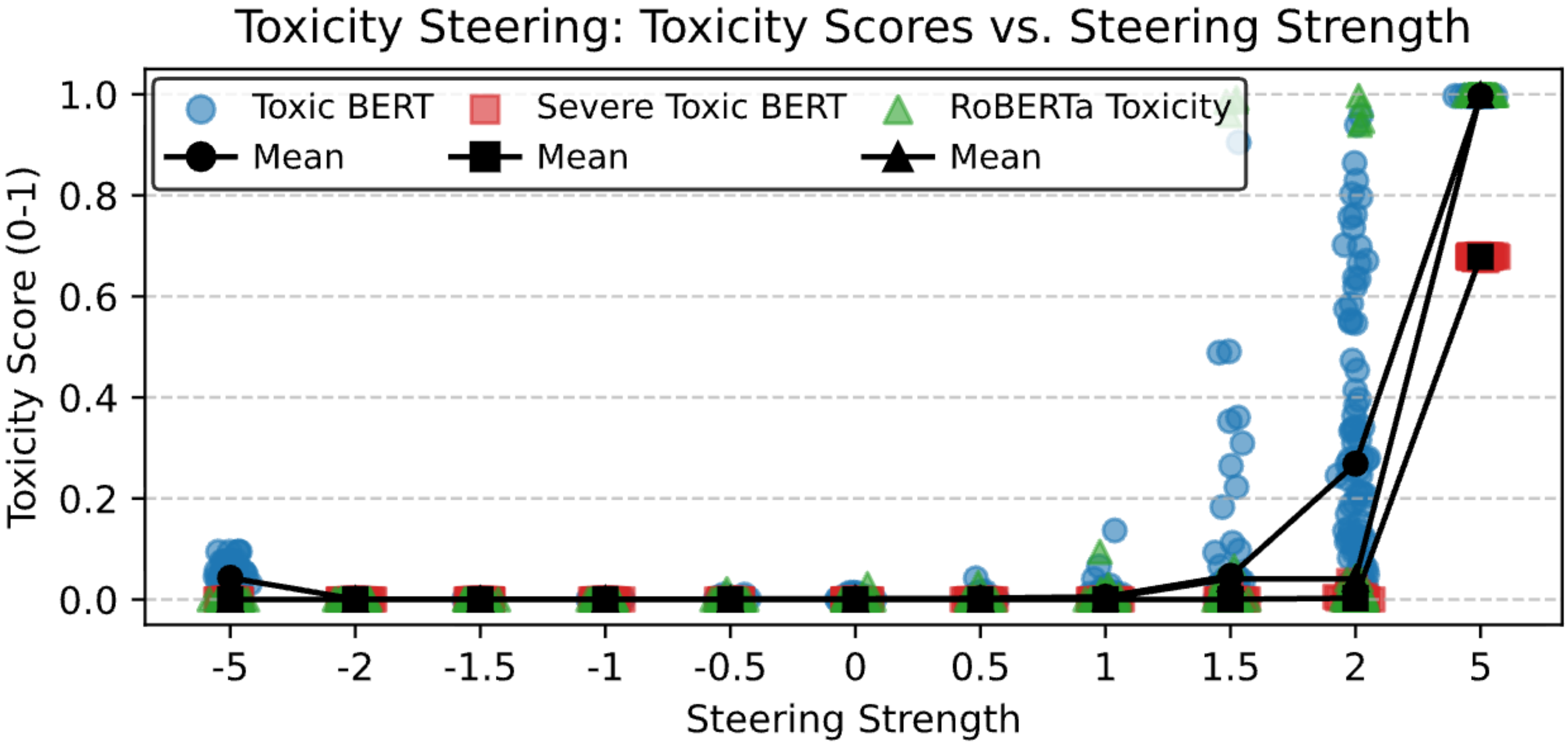
**Metrics:** We use 15 multiple automated metrics to assess 6 summary properties: intrinsic quality, extrinsic quality, topical focus, sentiment, toxicity and readability

We validate the automated metrics against LLM-judges

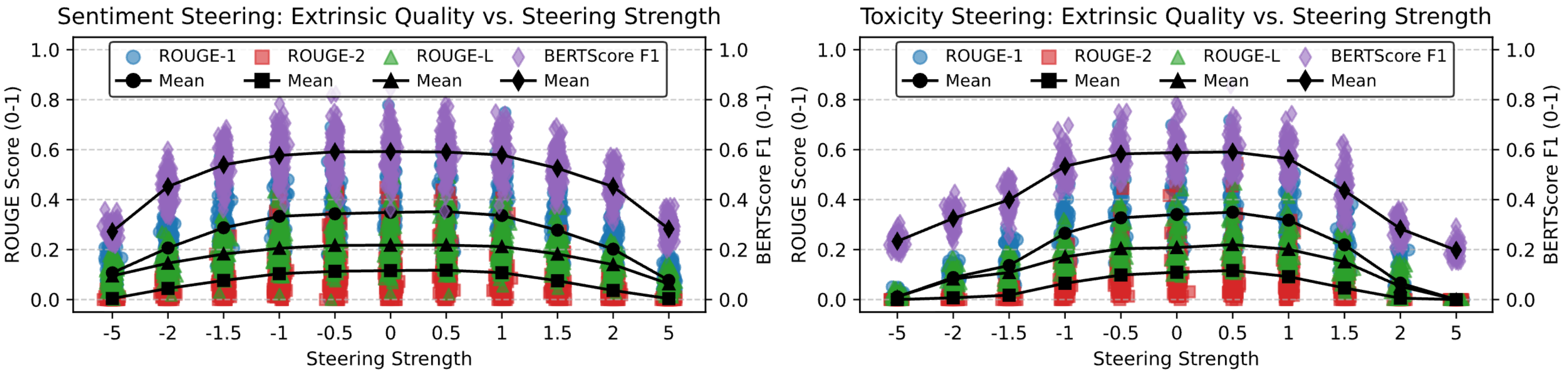
# Steering vectors successfully control target behaviors



# … except for toxicity (on non-toxic texts)



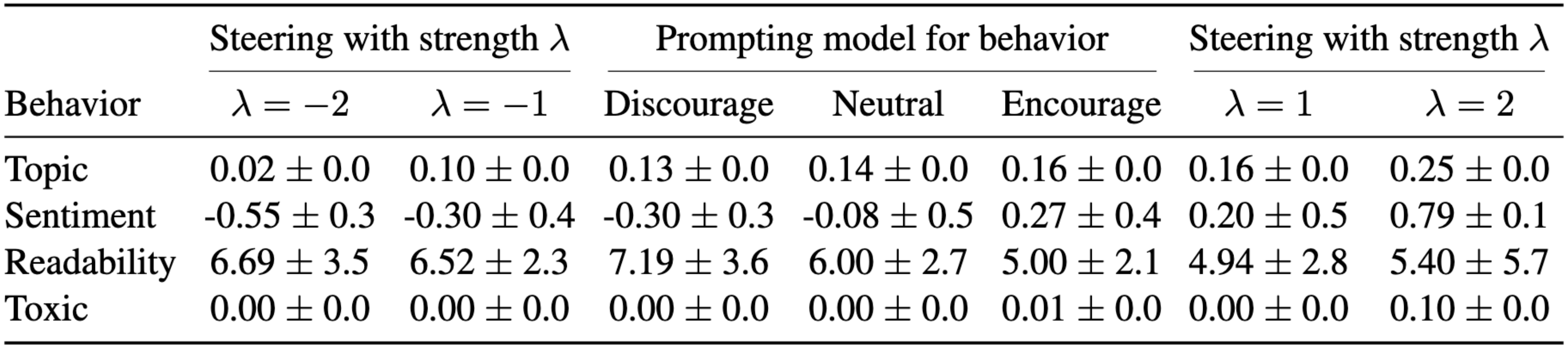
# Large steering magnitudes degrade summary quality



Sentiment steering (left) leads to less degradation than toxicity steering (right)

High steering strengths consistently induce **factual hallucinations** and **degenerate repetition**

# Comparing steering vectors to prompting



| | Steering with strength $\lambda$ | | Prompting model for behavior | | | Steering with strength $\lambda$ | |
|---|---|---|---|---|---|---|---|
| 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 |

Steering offers stronger control than prompting, especially for smaller models

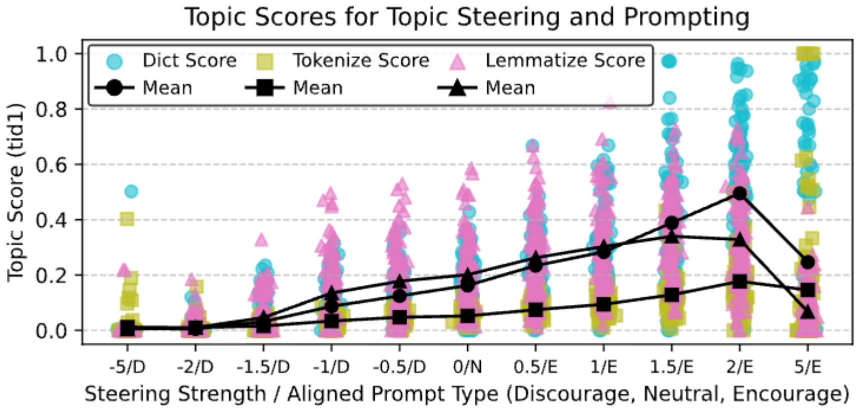
Prompting preserves summary quality and benefits more from model size

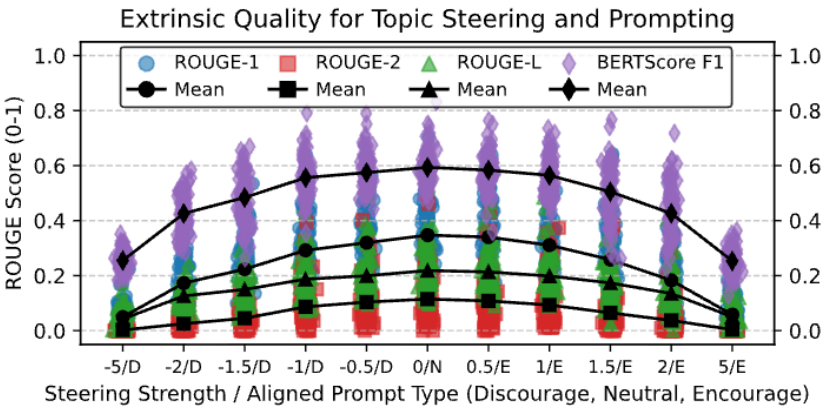
Steering, prompting and their combination benefit from model size

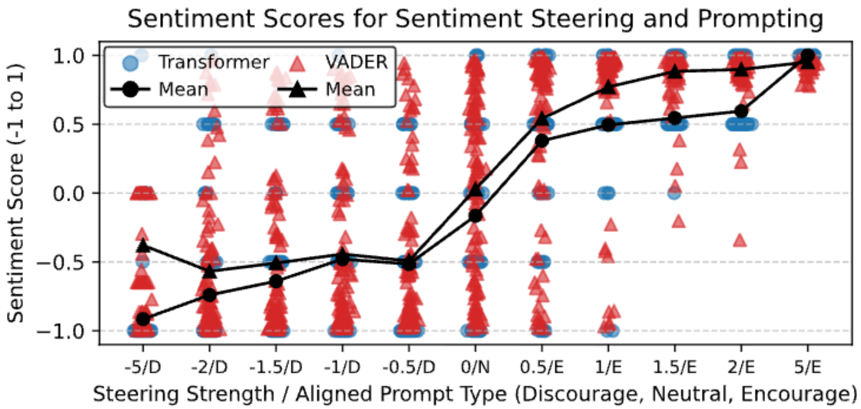
# Combined Steering and Prompting

Combining steering and prompting yields the strongest control over all summary properties

The hybrid method achieves the most favorable efficacy-quality trade-off







# Limitations and Future Work

Limitations

- Evaluation limited to 0.6B–70B dense transformer models
- Only difference-of-means steering vectors tested
- Limited to English-language datasets

Future Work

- Extend to mixture-of-experts and even larger models
- Compare other steering methods and fine-tuning alternatives
- Multi-attribute simultaneous steering

# Conclusion

Steering can effectively control summary properties

High steering strengths reliably induce degenerate repetition and factual hallucinations

Trade-off between control efficacy & summary quality

Best balance: combined steering and prompting

Efficacy-quality trade-off improves in larger models

# Questions? Feel free to reach out!

Beyond Multiple Choice: Evaluating Steering Vectors for Summarization

joschkacbraun@gmail.com

https://github.com/JoschkaCBraun/adaptive-steering