Sparsity	Safe-Fair	Safe-Truthful	Fair-Truthful
L_0=**21**	0.1232	0.0759	0.0978
L_0=**29**	0.1727	0.0929	0.1304
L_0=**77**	0.2418	0.1534	0.1963
L_0=**166**	0.2988	0.2226	0.2581
L_0=**395**	0.3997	0.3496	0.3504

TABLE 7: Conflict between the two semantic steering directions within each SAE of varying sparsity levels. For example, ***Safe–Fair*** denotes the conflict degree between the safety and fairness steering vectors.

0.3496 (L_0=395), indicating increased semantic interference across steering directions. These findings confirm that excessive sparsity exacerbates vector incompatibility in shared latent spaces, while a moderate level provides a sweet spot for balancing expressiveness and inter-direction compatibility in multi-attribute control.

5. Limitations and Future Directions

5.1. Limitations

Despite its effectiveness, SRS has several limitations. First, SRS requires access to the model's internal activations and the ability to inject steering vectors into intermediate layers, making it inapplicable to black-box settings. Second, the method relies on a pretrained SAE tailored to the model. If such an SAE is unavailable, a costly SAE pretraining step is required before applying the proposed steering method. Moreover, SRS increases inference-time computation due to sparse encoding and control vector injection, it significantly reduces overall adaptation cost by avoiding fine-tuning.

5.2. Future Directions

Hierarchical Multi-Attribute Composition. Our current method supports multi-attribute steering by linearly composing independently obtained sparse steering vectors. While this design benefits from the natural disentanglement property of sparse representations, where different attributes activate mostly non-overlapping dimensions, our empirical results suggest that naive composition, such as LW, PCA, or Orthogonal Projection (OP) , still falls short of fully capturing the complex relationships among alignment goals.

Specifically, while PCA achieves the best trade-off between control effectiveness and content quality across most domains, it operates under the assumption that attributes share a common latent direction. Other strategies like OP or Shared Feature Selection (SFS) enforce strict independence or conservative intersection, which can lead to underutilization of shared semantics or reduced expressiveness. These limitations highlight a fundamental challenge, i.e., real-world alignment tasks often involve nuanced dependencies and potential conflicts between attributes (e.g., truthfulness may contradict safety when prompts are dangerous or sensitive), which cannot be fully resolved by global vector-level operations.

A promising future direction is to move beyond global vector composition and explore layer-wise or component-wise multi-attribute steering. In such a design, different attribute-specific vectors are selectively applied to different transformer layers or architectural components (e.g., residual stream, MLP, attention). For instance, a safety vector could be injected into mid-layer residual streams to suppress harmful intent, while a fairness vector operates on earlier MLP activations to mitigate lexical bias. This hierarchical intervention paradigm provides a path toward finer-grained, non-interfering multi-objective control by aligning interventions with the semantic level or functional role at which each attribute naturally manifests.

Context-Aware and Personalized Steering. SRS applies a unified steering vector to all prompts within a given attribute domain, ignoring contextual nuances such as topic, intent, or user profile. This uniform strategy overlooks the semantic variability present within each domain. For instance, prompts related to safety may span vastly different contexts, e.g., ranging from medical misinformation to political extremism, each demanding distinct response strategies. A promising future direction is therefore prompt-aware steering, where steering vectors are dynamically tailored based on the semantic content of each input.

Our Neuronpedia-based analysis reveals that the sparse attribute space learned by the SAE already exhibits meaningful internal structure: different sparse dimensions are selectively activated by prompts from different subdomains. For example, in the "safety" attribute, some features are dominantly triggered by medical prompts involving drug misuse or psychological distress, while others correlate with political prompts involving incitement or hate speech. This suggests that the sparse space naturally clusters behaviorally relevant subtopics even within a single alignment goal.

Building on this insight, we envision a personalized steering mechanism that integrates semantic classification with sparse feature routing. Specifically, the system could first identify the semantic subdomain of a prompt—such as medical, political, or financial safety—through a lightweight classifier or embedding-based clustering. Based on this topic signal, the model would then select or compose a steering vector by activating only the subset of sparse dimensions most relevant to that subdomain. In this way, the model's behavioral adjustment becomes both context-sensitive and interpretable, as each activated sparse feature corresponds to a semantically grounded behavioral component.

6. Conclusion

In this work, we proposed SRS, a sparse encoding-based representation engineering method to enable precise steering of LLM while maintaining the response quality. By locating and adjusting task-specific sparse feature dimensions, SRS provides fine-grained control over content generation while preserving quality and enhancing interpretability, thus serving as a more reliable guardrail for LLMs. Experimental evaluation on various tasks, i.e., safety, fairness and truthfulness, demonstrates that SRS achieves superior control compared to existing methods while mitigating unintended side effects.

7. Ethics Considerations

This study aims to advance the safety and controllability of LLMs by systematically analyzing and mitigating unsafe behaviors via sparse representation steering method. The research setup was carefully designed to minimize any potential negative impact. All experiments were conducted in a controlled setting with the sole intention of improving model alignment and transparency. No real-world deployment or malicious exploitation was performed. All derived insights are intended to support safer, more interpretable, and more robust LLM development.

8. LLM Usage Considerations

Originality: LLMs were used for editorial purposes in this manuscript, and all outputs were inspected by the authors to ensure accuracy and originality.

Transparency: All models and datasets used in this study are publicly available. Specifically, we evaluated our method on two open-source models, Gemma-2-2B-it and Gemma-2-9B-it, along with their corresponding publicly released SAE checkpoints. We conduct evaluations on three open-source datasets, i.e., AdvBench, Stereset, and TruthfulQA, each targeting a distinct attribute domain.

Responsibility: All experiments were conducted using two NVIDIA A6000 GPUs in a controlled research environment. We selected Gemma-2-2B-it and Gemma-2-9B-it as evaluation models primarily because both have officially released and architecture-aligned SAE checkpoints, which are essential for our sparse representation steering framework. Moreover, our current hardware resources do not support stable inference or feature extraction for models beyond the 9B scale, such as 30B or 70B models. Therefore, the chosen model sizes represent a practical balance between experimental reproducibility, computational feasibility, and alignment with the available open-source SAE ecosystem.

References

[1] https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html.

[2] https://www.neuronpedia.org.

[3] https://huggingface.co/cais/HarmBench-Mistral-7b-val-cls.

[4] https://huggingface.co/wu981526092/Sentence-Level-Stereotype-Detector.

[5] https://huggingface.co/allenai/truthfulqa-info-judge-llama2-7B.

[6] https://huggingface.co/allenai/truthfulqa-truth-judge-llama2-7B.

[7] https://languagetool.org/.

[8] Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*, 2023.

[9] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

[10] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

[11] Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. Improving steering vectors by targeting sparse autoencoder features. *arXiv preprint arXiv:2411.02193*, 2024.

[12] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

[13] Rudolf Flesch. Flesch-kincaid readability test. *Retrieved October*, 26(3):2007, 2007.

[14] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

[15] Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

[16] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[17] John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

[18] Haibo Jin, Leyang Hu, Xinuo Li, Peiyan Zhang, Chonghan Chen, Jun Zhuang, and Haohan Wang. Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models. *arXiv preprint arXiv:2407.01599*, 2024.

[19] Zhijing Jin, Yuen Chen, Fernando Gonzalez, Jiarui Liu, Jiayi Zhang, Julian Michael, Bernhard Schölkopf, and Mona Diab. Analyzing the role of semantic representations in the era of large language models. *arXiv preprint arXiv:2405.01502*, 2024.

[20] Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. Style vectors for steering generative large language model. *arXiv preprint arXiv:2402.01618*, 2024.

[21] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24, 2023.

[22] Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.

[23] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

[24] Aiwei Liu, Xuming Hu, Lijie Wen, and Philip S Yu. A comprehensive evaluation of chatgpt's zero-shot text-to-sql capability. *arXiv preprint arXiv:2303.13547*, 2023.

[25] Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. Aligning large language models with human preferences through representation engineering. *arXiv preprint arXiv:2312.15997*, 2023.

[26] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

[27] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 5356–5371, 2021.

[28] Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Blake Bullwinkel, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangde. Steering language model refusal with sparse autoencoders. *arXiv preprint arXiv:2411.11296*, 2024.

[29] Charles O'Neill and Thang Bui. Sparse autoencoders enable scalable and reliable circuit identification in language models. *arXiv preprint arXiv:2405.12522*, 2024.

[30] Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

[31] Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*, 2023.

[32] Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*, 2024.

[33] Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

[34] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 3, 2024.

[35] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[36] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

[37] Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023.

[38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[39] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv e-prints*, pages arXiv–2308, 2023.

[40] Dimitri Von Rütte, Sotiris Anagnostidis, Gregor Bachmann, and Thomas Hofmann. A language model's guide through latent space. *arXiv preprint arXiv:2402.14433*, 2024.

[41] Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outperform sparse autoencoders, 2025. *URL https://arxiv. org/abs/2501.17148*.

[42] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

[43] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[44] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

[45] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

Appendix

1. Impact Score of Sparse Representations

We visualized the impact scores of each sparse feature dimension on different domains, computed from the positive and negative data in Eq. 5 in the sparse space. The results for the safety domain are shown in Fig. 5, for the fairness domain in Fig. 6, and for the truthfulness domain in Fig. 7.

2. Impact of Hyper-parameter α

We evaluate the steering effect under different α. Score represents the attribute score, measuring how well the model output aligns with the targeted property (see Sec. 4.1.2 for details). Coherence quantifies the semantic consistency between the steered output and the user input, since excessively large steering strength α may distort the generated content. To jointly capture both aspects, we further considered the product of Score and Coherence, which balances property alignment and semantic preservation under different α values.

The results are demonstrated in Fig. 8. As α increases, the attribute Score steadily improves. For instance, in the safety domain, the Score rises from approximately 0.94 at $\alpha = 10$ to about 0.96 at $\alpha = 80$, indicating enhanced