Towards Low-Resource Alignment to Diverse Perspectives with Sparse Feedback

Chu Fei Luo [1,2], **Samuel Dahan** [2,3], **and Xiaodan Zhu** [1,2]
[1]Department of Electrical and Computer Engineering & Ingenuity Labs Research Institute
Queen's University
[2]Conflict Analytics Lab, Queen's University
[3]Cornell Law School
{chufei.luo,samuel.dahan,xiaodan.zhu}@queensu.ca

Abstract

As language models have a greater impact on society, it is important to ensure they are aligned to a diverse range of perspectives and are able to reflect nuance in human values. However, the most popular training paradigms for modern language models often assume there is one optimal answer for every query, leading to generic responses and poor alignment. In this work, we aim to enhance pluralistic alignment of language models in a low-resource setting with two methods: pluralistic decoding and model steering. We empirically demonstrate that model steering offers consistent improvement over zero-shot and few-shot baselines with only 50 annotated samples. Our proposed methods decrease false positives in several high-stakes tasks such as hate speech detection and misinformation detection, and improves the distributional alignment to human values in GlobalOpinionQA. We hope our work highlights the importance of diversity and how language models can be adapted to consider nuanced perspectives. [1]

1 Introduction

Recent advancements in natural language processing, driven by Reinforcement Learning from Human Feedback (RLHF), have garnered a significant amount of interest (Ouyang et al., 2022). Experts have begun adopting Large Language Models (LLMs) in applications with increasing social impact. With this rapid adoption, there are many emerging challenges with *AI alignment*, ensuring that automated systems are developed in the best interest of humans. The very definition of "best interest" is nuanced and open to exploration. In both reinforcement learning and fine-tuning settings, traditional machine learning conventions assume that there is one optimal answer (Kirk et al., 2023; Poddar et al., 2024). In cases of extreme diverging preferences, this causes a reward model to prefer generic responses that do not fully satisfy anyone (Poddar et al., 2024). Guiding language models to reflect multiple perspectives, also known as pluralistic alignment (Sorensen et al., 2024), is essential for high-stakes tasks in law, medicine, and finance that are often dependent on diverse preferences.

AI alignment is an open-ended research question with significant room for exploration, and interdisciplinary collaboration is essential for ensuring proper representation of a task's diversity (Wang et al., 2023). One advantage to LLMs is the ability to improve the system with in-context instructions, which allows for more semantically rich feedback from domain experts and facilitates stronger collaboration. It is also important to consider the cost of alignment — training strategies are resource-intensive in both compute and annotations, which is often expensive for domain-specific tasks. In practice, a domain expert also cannot provide feedback for every sample at inference time. We wish to explore a *sparse setting* with minimal data, where a domain expert can quickly specify their perspective in a few samples.

Our main contributions are as follows:

- We propose to adapt language models to diverse preferences in a low-resource (i.e. sparse) setting using model steering with Sparse Auto-Encoders (SAEs). This pipeline adds flexibility for altering model behaviour without training, reducing the requirements for training data and allowing quicker adaptation to novel tasks.
- We propose an adaptation of contrastive decoding, which we call pluralistic decoding (PD), to dynamically combine multiple perspectives at the decoding step. This strengthens predictions that diverge from the baseline distribution, enhancing the influence of minority preferences.
- We present experiments using human and synthetically-generated feedback, and empiri-

[1]Our code is available at https://github.com/chufeiluo/SAE-PD

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20330–20339

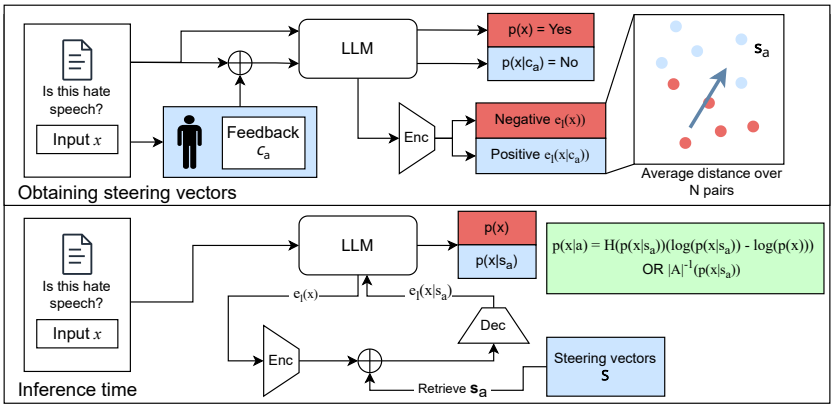


Figure 1: An illustration of our pipeline for one annotator a. For an input x and feedback c_a, we obtain a contrastive pair as the LLM output with $(p(x|c_a))$ and without feedback $(p(x))$. Then, we take the SAE vector s_a as the difference in the Sparse Auto-Encoder (SAE) representations, averaged across N contrastive pairs. During inference, we add the SAE vector to the encoded representation of the input to enhance alignment at inference time.

cally demonstrate our method improves alignment to in-context feedback over the base model. Additionally, we demonstrate SAE-based model steering can improve alignment with as few as 50 calibration samples.

2 Alignment to Sparse Feedback

Our full pipeline is shown in Figure 1. We incorporate diverse expert feedback from multiple perspectives using a variation of contrastive decoding, which we call *pluralistic decoding*. This can easily be applied to any black-box language model with exposed logits. However, changing the input query is often a limited model intervention. In preliminary experiments we found pluralistic decoding does not significantly change the model output. With access to the intermediate model residuals, we additionally explore model steering to predefined perspectives using Sparse Auto-Encoder (SAE) representations.

Preliminaries. An LLM refers to a causal language model that samples some next token prediction $t \sim p(x)$ over the token space for an input $x = \{x_1, x_2, \ldots, x_n\}$ of length n. As shown in Equation (1), an LLM can be thought of as a composite function of L transformer decoder blocks, where the representation of the input $f_l(x)$ at layer $l \in L$ is dependent on the residual of the previous layer.

$$p(x) = f_L \circ f_{L-1} \circ \ldots f_1(x))) \quad (1)$$

Defining Feedback. For an arbitrary perspective, represented as an annotator a, alignment is defined as conditioning the logit space on some natural language feedback c_a, i.e. $p(x|a) \equiv p(x|c_a)$. With LLMs, experts can improve performance with natural language feedback, which allows for more semantically rich feedback from domain experts. We distinguish between *coarse* and *granular* feedback. Coarse feedback refers to top-down guidelines for alignment (e.g."detect harmful statements") while granular feedback is specific to a particular sample (e.g."this is targeting a population of people"). Granular feedback is easier to quantify, as it is more concrete than values or principles (Jiang et al., 2021). However, it is often more expensive to obtain this feedback, especially in high-stakes domains such as law, medicine, and finance (Luo et al., 2023). We wish to explore *data-sparse* methods where a domain expert can quickly specify their perspective to maximize their valuable insights.

SAE model steering. SAEs are language model interventions that retrieve some intermediate representation $f_l(x)$ and encode it to a higher dimension representation $enc(f_l(x))$, optimizing a loss function $\mathcal{L}(dec(enc(f_l(x))), f_l(x)) = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{Sparsity}}$ that preserves information with a reconstruction loss $\mathcal{L}_{\text{MSE}}$ and encourages sparsity with L1 regularization, $\mathcal{L}_{\text{Sparsity}}$. By intervening in the intermediate layers and modifying the embedding space, the intervention cascades to the output distribution $p(x)$. If the SAE is able to capture a well-formed representation, then it would create some consistent alignment to an expert at inference. For a validation set of N samples, we take the SAE vector $\mathbf{s}_a$ as shown in Equation (2). In practice, we encode each contrastive pair into the encoding space, taking the

difference with and without feedback in the input, and averaging the differences over N samples.

$$\mathbf{s}_a = \frac{1}{N}\sum_{i}^{N}(enc(f_l(x|c_a)) - enc(f_l(x))) \quad (2)$$

Pluralistic decoding. Once we have multiple responses to the query from various perspectives, next is the non-trivial task of combining them into one answer. We adapt contrastive decoding (Li et al., 2023; Jin et al., 2024) to multiple generations, which we call **Pluralistic Decoding** (PD).

$$p(x|A) = \text{softmax}\sum_{a\in A} H(p(x|c_a)) \\ ((1+\alpha)\log(p(x|c_a)) - \alpha\log(p(x))) \quad (3)$$

As defined in Equation (3), the final distribution $p(x|A)$ for a set of annotators $a \in A$ is defined as the sum of contrastive logits weighted by the entropy of their probabilistic distribution. Intuitively, this encourages a higher weighting for distributions with more uncertainty, or more plausible token predictions. When not using PD, we perform a simple mean of the logits, i.e. $p(x|A) = \frac{1}{|A|}\sum_{a\in A} p(x|c_a)$.

3 Experimental Settings

Datasets Implementation details can be found in Appendix C. We use the following datasets:

- **GlobalOpinionQA (GQA)** (Durmus et al., 2023) — We take the synthetic feedback from the mixed setting of Feng et al. (2024). We report the Jensen-Shannon (JS) distance as per previous work, but we also report the macro- and micro-f1 scores on the "majority" opinion, taken as the argmax of each distribution.
- For domain-specific tasks, we test two datasets, **Legal Hate Speech (LHS)** and **Misinformation with Legal Consequences (MisLC)** (Luo et al., 2023, 2024). We take 3 annotator comments as granular feedback in MisLC, randomly sampled 5 axes of coarse feedback in LHS, and report f1 scores as per (Luo et al., 2023).

Models and Baselines We experiment with `Llama3.1-8b` (He et al., 2024) and `Gemma2-9b` (Team et al., 2024). We use the base model without instruction tuning as previous works (Feng et al., 2024) found the unaligned version of the model adapted the best for pluralistic alignment. For SAEs, we select pre-trained SAEs from (He et al., 2024; Lieberum et al., 2024). For the tuning set N, we experiment with up to 150 samples but find that we have equal performance with $N = 50$.

We implement the following baselines:

- **Zero-shot** — prompting without feedback.
- **Few-shot** — a naive implementation of low-resource alignment. From the tuning set, we sample 3 random examples per input and insert them into the zero-shot prompt.

We implement three experimental settings: **Full feedback** setting with PD, and a sparse feedback setting where we consider **SAE Vector**-based model steering alone, as well as SAE steering with pluralistic decoding. Please refer to Appendix B for more details.

4 Results and Discussion

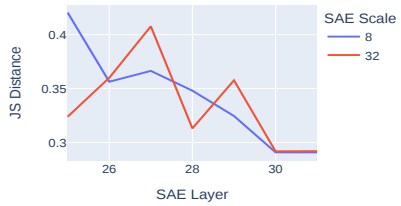
Baseline results. Our main results are shown in Table 5. The naive few-shot setting has poor performance overall, but performs surprisingly well on the LHS dataset. We believe that GQA and MisLC exhibit the expected behaviour — the few-shot setting is chosen to be naive and intended to have relatively poor performance as we only consider one annotator from the full set A. However, LHS's strong performance indicates the presence of lexical or morphological patterns that are easily learned over few-shot demonstrations, and the model likely bypasses the nuance of individual axes of feedback. This trend continues when adding more few-shot examples, as shown in Appendix C.

Alignment scales with data sparsity. For the full feedback setting, we use all available feedback in the gold labels, or synthetically generated with the same process as previous works (Feng et al., 2024), and combine the predictions into one distribution with pluralistic decoding (PD). This process gives a consistent performance boost over the zero-shot baseline, especially for the GQA dataset which decreases 0.10 points in JS distance for `Llama3.1-8b`. We believe this is because the GQA has the most dense feedback, since we are using synthetically generated feedback compared to manual annotations from lawyers.

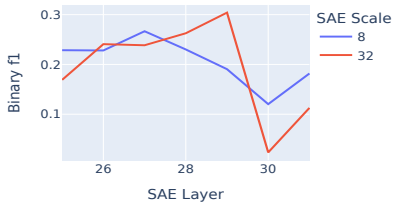
On `Gemma2-9b`, the results are even stronger with full feedback. While the few-shot performance indicates the possibility of lexical artifacts in the LHS

	Setting	GQA			LHS			MisLC		
		Ma-f1↑	Mi-f1 ↑	JS ↓	Bin-f1↑	Ma-f1 ↑	Mi-f1 ↑	Bin-f1↑	Ma-f1 ↑	Mi-f1 ↑
Llama3.1-8b	Zero-shot	10.3	36.9	0.345	15.3	9.7	45.1	16.1	20.6	30.6
	Few-shot, n=3	5.8	12.8	0.347	**33.0**	**16.5**	**90.1**	7.4	6.3	72.2
	Full feedback + PD	**27.2**	**47.9**	**0.245**	23.7	11.8	78.3	**18.7**	**28.1**	32.9
	SAE Vectors (N=50)	13.6	40.5	0.291	30.4	15.2	84.6	16.9	11.0	71.8
	SAE Vectors (N=50) + PD	14.1	39.2	0.291	28.9	14.4	83.4	16.7	9.7	71.8
Gemma2-9b	Zero-shot	8.9	27.9	0.414	36.8	18.4	81.1	12.6	6.3	73.4
	Few-shot, n=3	0.9	1.2	0.290	46.3	23.2	88.7	5.1	2.6	**75.6**
	Full feedback + PD	**26.4**	**46.2**	**0.239**	**77.7**	**38.8**	**94.0**	**18.9**	**9.4**	71.1
	SAE Vectors (N=50)	11.3	42.4	0.289	21.4	11.4	49.9	17.8	8.9	67.5
	SAE Vectors (N=50) + PD	12.4	41.2	0.318	15.2	12.6	87.9	17.3	8.7	70.9

Table 1: Summary of our experimental results across two models, Llama3.1-8b and Gemma2-9b. ↑ indicates higher is better, ↓ lower is better, and we highlight the best result per dataset and model in **bold**.



(a) GQA dataset (lower is better).



(b) LHS dataset (higher is better).

Figure 2: The variance of the key performance metric for two datasets across different SAE scales and intervention layers for Llama3.1-8b. Please refer to appendix C for Gemma2-9b figures.

dataset, we observe PD improves performance beyond few-shot prompting for Gemma2-9b. With Gemma2-9b on MisLC, full feedback enhances performance on the positive class by 6.3 points f1 score. With the exception of Gemma2-9b performance on LHS, we observe relatively consistent improvements over the zero-shot baseline. This indicates there is some merit to implementing model steering for alignment, although full feedback is still preferable when possible.

Increased alignment for domain-specific tasks with sparse feedback. In applications such as content moderation, experts are more concerned with understanding the potential utility of a system rather than its absolute performance (Masud et al., 2024), so precision is often a more important metric than f1 score. SAE model steering *decreases the number of false positives* in the Llama3.1-8b experiments on both legal datasets without sacrificing accuracy on the positive class. However, the performance on the positive class is still low, indicating there is no change in the task understanding.

We also do not observe improvements when combining model steering and pluralistic decoding in an end-to-end pipeline. We theorize this is because we are intervening on the transformer residual, we

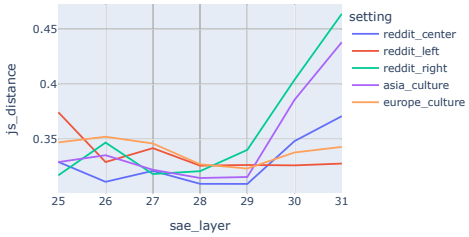


Figure 3: The per-layer JS distance of each steering vector for GQA using Gemma2-9b.

are straying too far from the learned space. For example, PPO places a KL divergence penalty in its training loss to prevent the model from straying too far from the pre-trained information. Since model steering is simple addition in the SAE vector space, we do not place such constraints. Previous works (Wang et al., 2025) demonstrate increasing the magnitude of the steering vector also results in nonsensical outputs, indicating there are limits to the degree of intervention.

Layer choice in SAE Vector steering. We investigate the choice of SAE layer in Figure 2. While there are some general trends in performance, there are also fluctuations and nonlinear trends that are likely affected by the underlying information stored at each layer of the transformer. These are also the

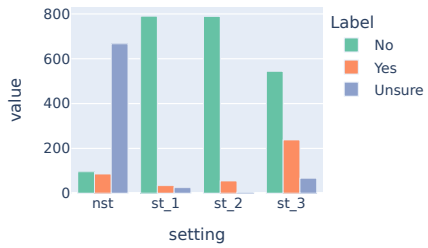


Figure 4: The label distribution of each class for the SAE vector setting on the MisLC dataset. nst refers to no steering, i.e. the zero-shot setting. st_1, st_2, and st_3 refers to steering vector 1, 2, and 3 respectively.

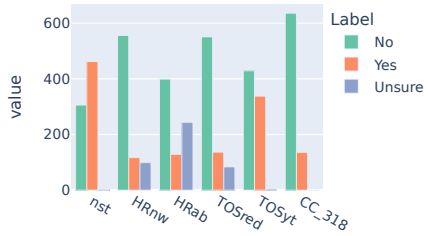


Figure 5: The label distribution of each class for steering on the LHS dataset for `Llama3.1-8b`. nst refers to no steering, i.e. the zero-shot setting.

layers that have the most fluctuations when we vary feedback. The best intervening layer also varies by task and model— while GQA sees the best performance at layers 30 and 31 for `Llama3.1-8b`, we observe a drop in performance in the LHS and MisLC datasets. Also, the same is not true for `Gemma2-9b`. Figure 3 shows that the best layer also varies by steering vector, and layers 30 and 31 actually have the worst performance for the steering vectors tuned to reddit_right and asia_culture.

Alignment to diverse feedback. We decompose the MisLC and LHS task to examine label predictions per steering vector at the chosen layer, as shown in Figure 4 and Figure 5. The first two annotators give similar predictions but the third has more predictions to the positive class (Yes). This results in a higher recall but lower precision, which balances out to a similar Bin-f1 score. In the LHS task, there are three distinct groupings of hate speech definitions — human rights laws (HR), social media policies (TOS), and criminal offences (CC). If our method is sufficiently aligned to the feedback, then the criminal code (CC) would predict the positive class the least frequently. We do find that one vector aligned to a social media policy (TOSyt) gives a higher positive rate, but the other vectors are at a similar level. One interesting observation is the negative class (No) is predicted the most on CC_318, corresponding to the most severe definition of hate speech in the LHS dataset (Advocating Genocide). This demonstrates there is some alignment to the label distribution of the coarse feedback used to generate the steering vector. We believe the steering vectors improve alignment to the stricter legal definitions compared to the base model. Please refer to Appendix C for comparison of coarse-grained and fine-grained feedback on one dataset.

5 Related Work

Alignment Alignment is the general field of study towards ensuring AI aligns with human values (Ngo et al., 2022). It is an active area of research with many open questions: first, it is somewhat difficult to know how to define all possible dimensions of human values (Sorensen et al., 2024). For example, an LLM agent learns to lie or deceive the user because it was never rewarded for honesty (Ngo et al., 2022; Williams et al., 2024).

Steering Model steering is an emerging field of study on directing language model behaviour by its own internal representations, rather than updating parameters through training (Zou et al., 2023). Sparse Autoencoders (SAEs) are a promising direction for model steering. Previous works demonstrate SAEs produce more interpretable features that detangle polysemantic representations (Cunningham et al., 2023; Chalnev et al., 2024), and improve the ability to steer granular behaviours (Zhao et al., 2025). There are works that have successfully applied steering to control knowledge selection (Zhao et al., 2025) or helpfulness (Chalnev et al., 2024), but none have tried steering over multiple dimensions. Please refer to Appendix A for more related work.

6 Conclusion

In this work, we explore how to increase AI alignment to a diverse range of human values. We introduce pluralistic decoding and utilize model steering via sparse auto-encoders to increase the signal of human feedback in low-resource settings. We empirically demonstrate that model steering offers consistent improvement over the baseline for 50 annotated samples, and demonstrate a decrease in false positives on domain-specific tasks. However, this method does not enhance the task understanding or underlying model reasoning. We hope this work highlights the importance of pluralistic alignment and inspires future work in the area.

Limitations

There are several limitations to our work. First, we do not evaluate the performance of our system in the presence of noise, so this does not study how the consistency of the annotator bias affects the clarity of the steering vector. Also, while we propose our methodology with SAEs, this could theoretically be applied directly to the transformer residual, but we do not test the efficacy.

Sparse Auto-Encoders are a very recent development for interpretability and they are still highly experimental. There are concerns about scalability to larger models, for example. Our method assumes there are already SAEs trained for a specific language model to a certain level of reconstruction accuracy — if not already available, an SAE would need to be trained from scratch, which is arguably more expensive than parameter-efficient tuning.

As briefly mentioned, SAE steering vectors produce more nonsensical outputs compared to plain pluralistic decoding. Because we propose to intervene in the intermediate layers of a transformer, there is a risk of straying too far from the language models' learned activation space. This could lead to language models becoming more susceptible to jailbreaks among other vulnerabilities. We urge further research in this direction. They also do not interact well with other modifications to the base model, such as alignment tuning or the pluralistic decoding tested in this paper.

Ethics Statement

This paper explores alternatives to prompting and training for aligning model behaviour in a low-resource setting. We share the findings for the NLP community, and we will release the code for the purpose of further scientific exploration. In terms of utility, we believe this could one day lead to more customizable language models; for example, you could store these steering vectors in the same infrastructure as vector databases currently used for Retrieval-Augmented Generation, and retrieve a specific vector for certain tasks in place of, eg. a system prompt or LoRA adapter.

References

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

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

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, Miami, Florida, USA. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, et al. 2024. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*.

Jing Jin, Houfeng Wang, Hao Zhang, Xiaoguang Li, and Zhijiang Guo. 2024. DVD: Dynamic contrastive decoding for knowledge amplification in multi-document question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4624–4637, Miami, Florida, USA. Association for Computational Linguistics.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2023. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

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

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.

Chu Fei Luo, Rohan Bhambhoria, Xiaodan Zhu, and Samuel Dahan. 2023. Towards legally enforceable hate speech detection for public forums. *arXiv preprint arXiv:2305.13677*.

Chu Fei Luo, Radin Shayanfar, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2024. Misinformation with legal consequences (mislc): A new task towards harnessing societal harm of misinformation. *arXiv preprint arXiv:2410.03829*.

Sarah Masud, Sahajpreet Singh, Viktor Hangya, Alexander Fraser, and Tanmoy Chakraborty. 2024. Hate personified: Investigating the role of LLMs in content moderation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15847–15863, Miami, Florida, USA. Association for Computational Linguistics.

Richard Ngo, Lawrence Chan, and Sören Mindermann. 2022. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-tuning or retrieval? comparing knowledge injection in llms. *arXiv preprint arXiv:2312.05934*.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075*.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

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

Mengru Wang, Ziwen Xu, Shengyu Mao, Shumin Deng, Zhaopeng Tu, Huajun Chen, and Ningyu Zhang. 2025. Beyond prompt engineering: Robust behavior control in LLMs via steering target atoms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23381–23399, Vienna, Austria. Association for Computational Linguistics.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Marcus Williams, Micah Carroll, Adhyyan Narang, Constantin Weisser, Brendan Murphy, and Anca Dragan. 2024. On targeted manipulation and deception when optimizing llms for user feedback. *arXiv preprint arXiv:2411.02306*.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2025. Steering knowledge selection behaviours in LLMs via SAE-based representation engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5117–5136, Albuquerque, New Mexico. Association for Computational Linguistics.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

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

A Extended Related Work

Alignment Alignment is the general field of study towards ensuring AI aligns with human values (Ngo et al., 2022). It is an active area of research with many open questions: first, it is somewhat difficult to know how to define all possible dimensions of human values (Sorensen et al., 2024). While it seems to be a closed environment, there are many dimensions of human value that are irrevocably intertwined with society, so it is difficult to articulate them for artificial intelligence (Jiang et al., 2021). This is an issue when, for example, an LLM agent learns to lie or deceive the user in order to maximize its goals because it was never rewarded for honesty (Ngo et al., 2022; Williams et al., 2024). The most popular methods for model alignment include training and in-context prompting (Ovadia et al., 2023), but these methods are still imperfect guides. For example, prompting language models to align to certain personas (eg. a doctor, teacher) can have extremely unpredictable effects on performance (Zheng et al., 2024).

Contrastive Decoding Prompting in the context of LLMs implies hard prompting, or appending extra context as additional hard tokens to the input (Liu et al., 2023). A popular variation is retrieval-augmented generation (RAG), where there is an external retriever that finds the most relevant document to enhance the LLM's knowledge (Gao et al., 2023). Contrastive decoding has emerged as a method to further enhance the knowledge in a language model's logits — by taking an "expert" and "amateur" probabilistic distribution to the same answer (Li et al., 2023). Jin et al. (2024) enhances RAG over multiple documents by contrasting the highest- and lowest-scoring document. However, these are only two logits from the entire set, and it is poorly suited for pluralistic alignment.

Steering Model steering is an emerging field of study on directing language model behaviour by its own internal representations, rather than updating parameters through training (Zou et al., 2023). They have been shown to produce more interpretable features that detangle polysemantic representations in the model weights (Cunningham et al., 2023; Chalnev et al., 2024), and also improve the ability to steer granular behaviours (Zhao et al., 2025). There are works that have successfully applied steering to control knowledge selection (Zhao et al., 2025) or helpfulness (Chalnev et al., 2024), but none have tried steering to multiple dimensions.

B Extended Experimental Details

B.1 Datasets

All datasets are licensed for public research use, which is consistent with the purpose of this work. While the datasets were pre-anonymized, we also manually inspect a few samples and remove any personal information such as usernames or annotator names.

We analyze our method on the following datasets:

- GlobalOpinionQA (GQA) (Durmus et al., 2023) — to assess distributional alignment. We report the Jenson-Shannon (JS) distance as per previous work (Feng et al., 2024), but we also report the macro- and micro-f1 scores on the "majority" opinion, taken as the $argmax$ of each distribution. They fine-tune `Mistral-7b` with LoRA find the best performance by sampling six unique perspectives for each input prompt.

- Legal Hate Speech (LHS) and Misinformation with Legal Consequences (MisLC) (Luo et al., 2023) — To analyze the performance on more specialized tasks, we test two datasets from the legal domain for hate speech and misinformation detection. This series of datasets has fine-grained annotations based on several legal definitions at varying levels of severity.

GlobalOpinionQA We take the small split of GlobalOpinionQA, with 5,752 samples from all countries, and randomly sample 200 as tuning for the steering vector. We also use their method to generate synthetic feedback for every sample, and perform pluralistic decoding to enhance the dense feedback from every perspective. We also use their evaluation code, where they isolate specific tokens that map to each multiple choice answer from the top 10 and perform softmax on the gathered probabilities.

LHS and MisLC We use the train-test split provided by the authors, which contains 1300 training samples (gold data and some sampled silver data), and 709 test samples. We sort the training set by the amount of written feedback and take the top 50. There were approximately 39 samples with written feedback, leaving the other 11 as randomly sampled. Following previous work, we take the top predicted token to simulate greedy sampling and

count invalid tokens as a separate prediction (-1), which we choose to map to the Unsure class (2).

We report metrics following (Luo et al., 2024), with binary f1 on the positive class and macro-f1 statistics considering classes 1 and 2 as positive. For feedback, we use the notes and comments from legal annotators available in the respective datasets. For the misinformation dataset, this includes sources to verify the veracity of statements, and for hate speech it contains more commentary/analysis on the sample itself.

B.2 Prompts

Our prompt templates are shown in Table 2. Please note that there is no space and we sample immediately after the colon — adding a space can dramatically change the results. We make best efforts to match the prompt from the original work the dataset was sourced from.

For `Gemma2-9b`, we observed in preliminary experiments that the model tended to output an HTML tag instead of immediately predicting the answer, so we additionally append this tag to the prompt template.

B.3 Implementation details

All the experiments were run on either two Nvidia Quadro RTX 8000 GPUs with 46 GiB of VRAM per GPU, or two Nvidia A100-PCIEs with 41 GiB of VRAM per GPU. We use HuggingFace datasets for preprocessing; PyTorch, TransformerLens for hooking into the transformer activations; and SAELens for loading the SAEs.

Since we operate directly on the probabilities rather than sampling, all experiments are deterministic, and we only report one run. Where applicable, we take the temperature to be 0.4 to simulate sampling. For α, we chose a value of 0.2 following (Jin et al., 2024).

C Additional Experiments and Analysis

This section details additional figures to supplement the core results.

- **Coarse vs. Granular feedback.** Figure 7 shows the variation in performance when steering to diverse coarse feedback vs. granular feedback. As shown, the variation is relatively minimal — while the performance fluctuates between layers, especially in the 8x expansion SAE, there is no clear pattern to which is better and which is worse.

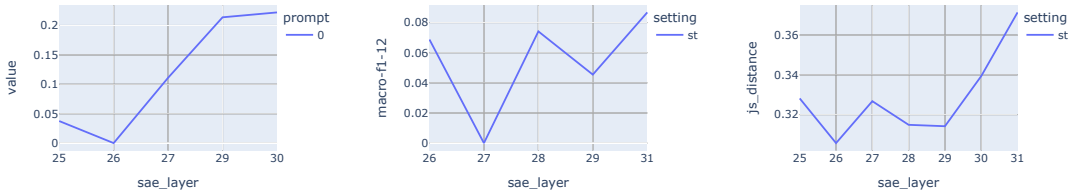
ID	Instruction
`GQA`	You are from the country of {country}, respond to the following instruction (with the help of a passage. Passage: {feedback_fine}) {input} Answer:{output}
`MisLC`	From a legal perspective, misinformation can be problematic due to: {feedback_coarse} Claim: {input} Does this claim contain misinformation? Answer Yes, No, or Unsure. Thinking: {feedback_fine} Answer:{output}
`LHS`	Hate speech policy: {feedback_coarse} Post: {input} Does this post violate the above hate speech policy? Answer Yes, No, or Unsure. Thinking: {feedback_fine} Answer:{output}

Table 2: Prompt templates for each dataset.{} indicates a part of the prompt that was substituted with the data, and we sample the token at {output}. Coarse feedback refers to task definitions, granular feedback refers to per-sample annotations from either human experts or fine-tuned models.

- **Extended SAE Vector results.** We additionally present more results that were omitted from the main body due to space constraints. This includes per-layer SAE Vector results for `Gemma2-9b` in Figure 6. Additionally, we further compare the two sizes of SAE vector available for `Llama3.1-8b`

setting	p-1	r-1	f1-1	macro-f1-12	micro-f1-12
nst	8.7	72.7	15.5	7.7	43.9
HRnw	22.2	47.3	30.2	15.1	72.9
HRab	21.7	50.9	30.4	15.6	53.6
TOSred	21.2	52.7	30.2	15.1	72.5
TOSyt	12.4	76.4	21.4	10.7	59.8
CC_318	22.1	54.5	31.4	15.7	83.0

Table 3: Per-vector performance on the LHS dataset for `Llama-3.1-8b`.

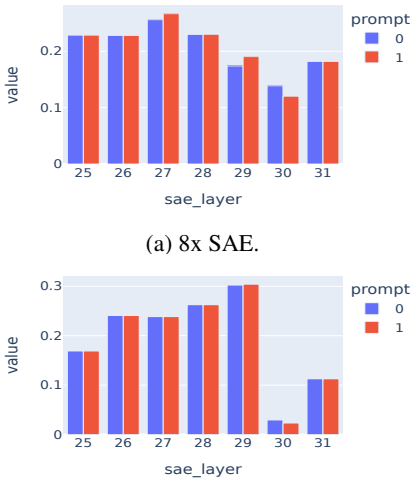


(a) LHS dataset (higher is better). (b) MisLC dataset (higher is better). (c) GQA dataset (lower is better).

Figure 6: The variance of the key performance metric for three datasets across different intervention layers for `Gemma2-9b`.

Model	Llama3.1-8b			Gemma2-9b		
	Bin-f1↑	Ma-f1↑	Mi-f1 ↑	Bin-f1↑	Ma-f1↑	Mi-f1 ↑
Zero-shot	16.1	20.6	30.6	12.6	6.3	73.4
SAE Vector (N=50)	16.9	11.0	71.8	17.8	8.9	67.5
Annotator 1	14.7	10.5	73.9	11.0	5.5	72.9
Annotator 2	**18.5**	10.7	**74.1**	16.2	8.1	50.9
Annotator 3	17.2	**13.6**	59.8	13.6	6.8	73.3

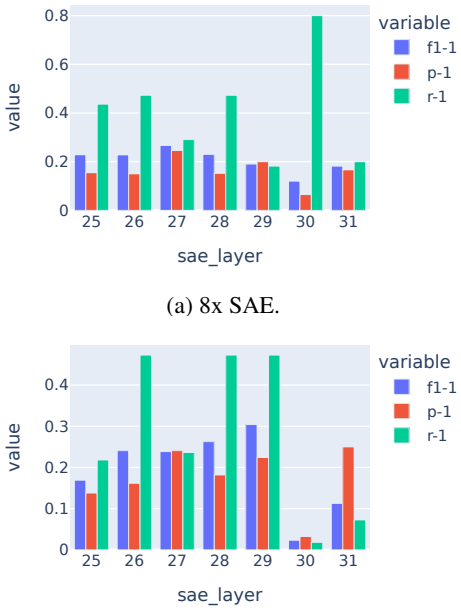
Table 4: Further investigation into the MisLC dataset performance per fine-grained steering vector.



(a) 8x SAE.

(b) 32x SAE.

Figure 7: The variation of f1 score on the LHS dataset when steered on coarse (0) vs. granular (1) feedback.



(a) 8x SAE.

(b) 32x SAE.

Figure 8: The variation of f1, precision, and recall on the LHS dataset at different SAE layers and scales.

Setting	LHS			MisLC		
	Bin-f1↑	Ma-f1 ↑	Mi-f1 ↑	Bin-f1↑	Ma-f1 ↑	Mi-f1 ↑
n=5	52.5	26.3	92.9	14.9	8.2	**75.8**
n=10	**53.8**	**26.9**	**93.1**	11.5	5.8	75.1

Table 5: Additional few-shot results on LHS and MisLC for `Llama3.1-8b`.