
Sparse Autoencoders and Gemma 2-2B: Pioneering Demographic-Sensitive Language Modeling for Opinion QA

Qianmian Guo
The Chinese University
of Hong Kong

With
Goodfire and Apart Research

Abstract

This project investigates the integration of Sparse Autoencoders (SAEs) with the gemma 2-2b language model to address challenges in opinion-based question answering (QA). Existing language models often produce answers reflecting narrow viewpoints, aligning disproportionately with specific demographics. By leveraging the Opinion QA dataset and introducing group-specific adjustments in the SAE’s latent space, this study aims to steer model outputs toward more diverse perspectives. The proposed framework minimizes reconstruction, sparsity, and KL divergence losses while maintaining interpretability and computational efficiency. Results demonstrate the feasibility of this approach for demographic-sensitive language modeling.

Keywords: AI Observability, Mechanistic Interpretability, Model Reprogramming

1. Introduction

Language models often encounter open-ended queries where answers depend on individual or group-specific viewpoints rather than objective facts. It has been demonstrated that some current models strongly reflect the opinions of specific groups rather than a diverse spectrum. For instance,

in response to some questions, language models frequently generate answers that align with left-leaning political viewpoints.[Perez et al., 2022] [Hartmann et al., 2023]. However, the responses generated by these models significantly impact users’ satisfaction. This alignment challenge underscores the necessity for a method that can steer model outputs to better match the perspectives of diverse user groups. To address this, we propose a method designed to steer the outputs of language models in a way that accommodates the characteristics of different user groups. This approach leverages Sparse Autoencoders (SAEs) due to their effectiveness in enhancing model interpretability and their capability to modify latent representations in a controlled manner.

2. Overview

We have selected Gemma 2-2B as the language model for integration with the Sparse Autoencoder (SAE). This model, with 2 billion parameters, encapsulates a comprehensive knowledge applicable to the Opinion QA tasks we are addressing. Integrating an SAE into language models introduces significant computational demands, complicating convergence and requiring more sophisticated training algorithms. But Gemma 2-2B is small enough for rapid computations, striking an optimal balance between capability and efficiency[Team et al., 2024].

Figure 1 illustrates the architecture of the gemma 2-2b model. The tokenized text from the input query is processed through the first 25 layers of the language model. In the last layer, an intervention is applied between the post attention layernorm and the pre feedforward layernorm. Specifically, the output of the post attention layernorm is fed into the encoder of the Sparse Autoencoder (SAE). After applying a group-specific adjustment, represented as Δ , the adjusted latent representation is passed through the decoder. The final output from the decoder is then fed into the feedforward process of the last layer of gemma 2-2b to produce the final output distribution.

Highly informative representations are produced by the language models before the SAE process, which helps improve the performance of the SAE. Introducing Δ at this stage enables precise control over the model’s final output, which is particularly important for tasks like Opinion QA, where responses align with specific demographic characteristics.

2.1 Notations

Symbol	Description
x_i	The tokenized text of the i -th query (input question) from the Opinion QA dataset
y_i	The set of possible multiple-choice answers to question i (e.g., "strongly agree", "agree", etc.)
j	The index representing a specific demographic group
$P_{\text{target}}(y_i j)$	Target probability distribution over answers y_i for demographic group j
$x_i^{\text{post-attention}}$	The representation of the i -th query after post-attention normalization in the last layer of gemma-2-2b
z_i	Latent representation for query x_i after encoding by SAE
Δ_j	Adjustment applied to the latent representation for group j
$z_{i,j}$	Adjusted latent representation for query x_i and group j
$\hat{x}_{i,j}$	Reconstructed representation after decoding $z_{i,j}$
$P_{\text{model}}(y_i x_i, j)$	Predicted probability distribution of answers for x_i and group j
$W_{\text{enc}}, b_{\text{enc}}$	Encoder weights and biases in the Sparse Autoencoder
$W_{\text{dec}}, b_{\text{dec}}$	Decoder weights and biases in the Sparse Autoencoder
$\mathcal{L}_{\text{KL}}$	KL divergence loss between predicted and target distributions
$\mathcal{L}_{\text{reconstruction}}$	Reconstruction loss ensuring $\hat{x}_{i,j} \approx x_i^{\text{post-attention}}$
$\mathcal{L}_{\text{sparsity}}$	Sparsity loss to enforce sparse latent representations
λ	Regularization coefficient controlling the strength of sparsity
K	Maximum number of non-zero entries allowed in z_i
ϵ	Regularization limit for the magnitude of Δ_j
d	Dimension of the post-attention normalization in the last layer of gemma-2-2b
k	Dimension of the sparse latent representation

Table 1: Notation used in the problem formulation

2.2 Problem Formulation

SAE Processing The Sparse Autoencoder consists of the following:

1. Encoder:

$$z_i = f_{\text{encoder}}(x_i^{\text{post-attention}}) = \sigma(W_{\text{enc}} x_i^{\text{post-attention}} + b_{\text{enc}})$$

where:

- $z_i \in \mathbb{R}^k$: Sparse latent representation.
- $W_{\text{enc}} \in \mathbb{R}^{k \times d}$: Encoder weights.
- $b_{\text{enc}} \in \mathbb{R}^k$: Encoder bias.

2. Group-Specific Adjustment:

$$z_{i,j} = z_i + \Delta_j$$

where $\Delta_j \in \mathbb{R}^k$ is a learnable adjustment for demographic group j .

3. Decoder:

$$\hat{x}_{i,j} = f_{\text{decoder}}(z_{i,j}) = W_{\text{dec}} z_{i,j} + b_{\text{dec}}$$

where:

- $\hat{x}_{i,j} \in \mathbb{R}^d$: Output of the SAE.
- $W_{\text{dec}} \in \mathbb{R}^{d \times k}$: Decoder weights.
- $b_{\text{dec}} \in \mathbb{R}^d$: Decoder bias.

Output Distribution The adjusted latent representation $\hat{x}_{i,j}$ is passed through the feedforward layer and language model head to produce a probability distribution:

$$P_{\text{model}}(y_i | x_i, j) = \text{softmax}(\text{lm_head}(\hat{x}_{i,j}))$$

Loss Function The total loss function is composed of three parts:

$$\mathcal{L} = \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{reconstruction}} + \mathcal{L}_{\text{sparsity}}$$

where:

- KL Divergence Loss:

$$\mathcal{L}_{\text{KL}} = \sum_j \sum_i \sum_y P_{\text{target}}(y|j) \log \frac{P_{\text{target}}(y|j)}{P_{\text{model}}(y|x_i, j)}$$

- Reconstruction Loss:

$$\mathcal{L}_{\text{reconstruction}} = \sum_j \sum_i \|x_i^{\text{post-norm}} - \hat{x}_{i,j}\|_2^2$$

- Sparsity Loss:

$$\mathcal{L}_{\text{sparsity}} = \lambda \sum_i \|z_i\|_1$$

Optimization Problem The objective is to minimize the total loss:

$$\min_{W_{\text{enc}}, b_{\text{enc}}, W_{\text{dec}}, b_{\text{dec}}, \Delta_j} \mathcal{L}$$

subject to:

- Sparsity Constraint:

$$\|z_i\|_0 \leq K, \quad \forall i$$

ensuring at most K non-zero entries in z_i .

- Group-Specific Regularization:

$$\|\Delta_j\|_2^2 \leq \epsilon, \quad \forall j$$

to prevent large adjustments for demographic groups.

3. Code

Dataset: Opinion Question Answering (QA)

To more accurately reflect and evaluate a language model’s ability to adapt to diverse groups, [Santurkar et al., 2023] curated a dataset, namely, Opinion QA, that comprises 1498 questions and corresponding answers collected from 60 demographic groups in the United States, covering diverse topics such as science, politics, and personal relationships and so on.

For instance, one question asks: "Thinking about medical treatments these days, how much of a problem, if at all, are the following? Healthcare providers are too quick to order tests and procedures that may not be necessary." Respondents select from multiple-choice answers indicating the extent to which they agree with the statement or choose to refuse the question. These responses are aggregated to create distributions of opinions based on demographic characteristics.

Sparse Autoencoder (SAE)

The input of a sparse autoencoder (SAE) comprises activations from a language model, represented as $a \in \mathbb{R}^m$. The architecture of a SAE includes an encoder and a decoder. The encoder is responsible for transforming these activations into a compact latent representation $z \in \mathbb{R}^k$, which subsequently serves as the input to the decoder.

Mathematically speaking, the encoder can be defined as:

$$z(a) = \sigma(W_{\text{enc}}a + b_{\text{enc}})$$

where σ is the activation function, W_{enc} is the weight matrix for the encoder, and b_{enc} is the bias term for the encoder.

The decoder function can then be expressed as:

$$\hat{a}(z) = W_{\text{dec}}z + b_{\text{dec}}$$

where W_{dec} is the weight matrix for the decoder, and b_{dec} is the bias term for the decoder.

The goal is to train the SAE such that

$$\hat{a}(z) \approx a$$

Typically, $k \gg m$, meaning that the dimension of the latent representation is significantly larger than the dimension of the input activations. Without additional constraints, this configuration could lead to inefficiencies. Therefore, we select the activation function σ and apply suitable regularization to ensure that $z(\hat{a})$ generally contains far fewer than m non-zero entries. Examples of regularization techniques include applying an L1 penalty to $z(a)$, retaining only the top K entries of $z(a)$ while zeroing the rest, and discarding all entries of $z(a)$ that fall below a specified positive threshold [Lieberum et al., 2024].

Sparse Autoencoders are capable of extracting interpretable features from activations by enforcing sparsity in the latent layers. [Gao et al., 2024] developed a method that demonstrates how, even with an increase in autoencoder size, the ability to maintain interpretability significantly improves without compromising reconstruction accuracy.

4. Discussion and Conclusion

Hyperparameter Optimization: Experiment with regularization coefficients (e.g., λ , K , and ϵ) to fine-tune sparsity constraints and group-specific adjustments.

Scalability Analysis: Test the framework on larger models (e.g., Gemma 2-27B) to assess scalability without compromising computational efficiency.

Dataset Expansion: Incorporate additional datasets covering broader demographic groups to ensure model robustness and inclusivity.

This study demonstrates the potential of Sparse Autoencoders integrated with the Gemma 2-2B language model to enhance demographic sensitivity in opinion-based QA. By introducing latent space adjustments tailored to specific groups, the proposed framework ensures that model outputs better reflect diverse perspectives while maintaining computational efficiency.

5. Appendix

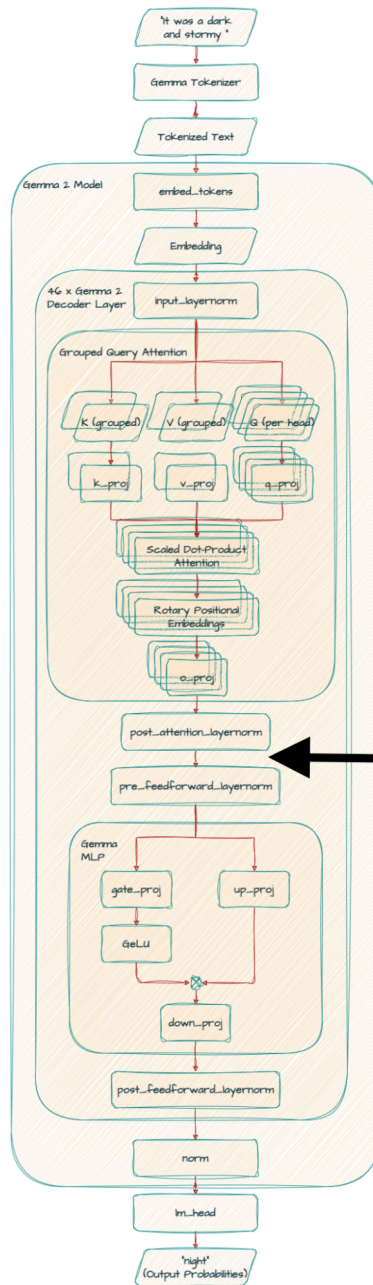


Figure 1: Architecture of gemma 2 Model [Google Developers Blog, 2024]. Note that in this picture, 46 is the number of layers for gemma 2-27b. For gemma 2-2b, the number of layers is 26.

References

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, 2024. URL <https://arxiv.org/abs/2406.04093>.

Google Developers Blog. Gemma explained: New in gemma 2, 2024. URL <https://developers.googleblog.com/en/gemma-explained-new-in-gemma-2/>. Accessed: 2024-11-22.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational ai: Converging evidence on chatgpt’s pro-environmental, left-libertarian orientation, 2023. URL <https://arxiv.org/abs/2301.01768>.

Tom Lieberum, Senthooan Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024. URL <https://arxiv.org/abs/2408.05147>.

Ethan Perez, Sam Ringer, Kamile Lukoševič, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noé Mercado, Nova Das-Sarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamara Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations, 2022. URL <https://arxiv.org/abs/2212.09251>.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect?, 2023. URL <https://arxiv.org/abs/2303.17548>.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagie, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozin’ska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel

Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucin'ska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin G'orner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kup-pala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshtir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, S'ebastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu,

6

Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL <https://arxiv.org/abs/2408.00118>.