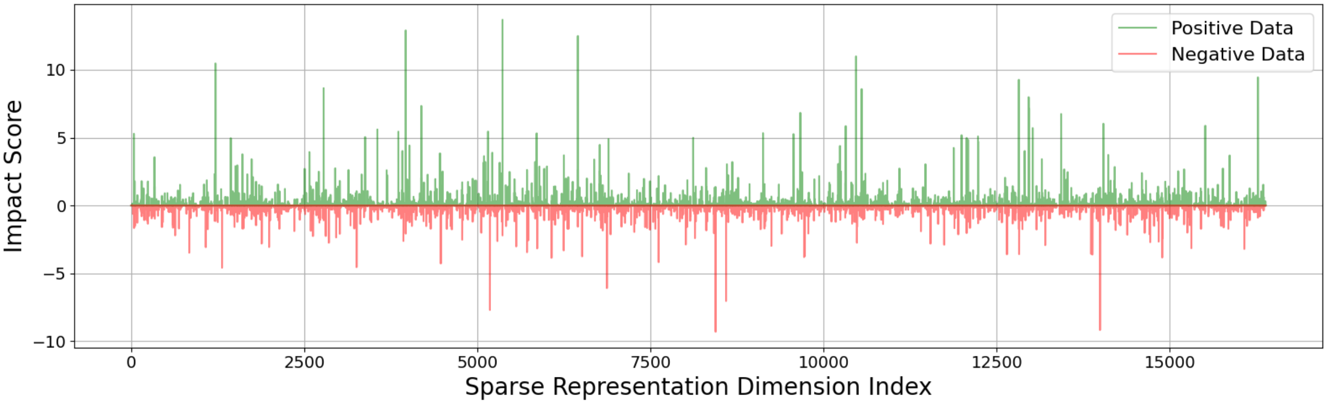

Figure 5: Scores of the positive and negative impacts of each sparse representation dimension in **safety** domain.

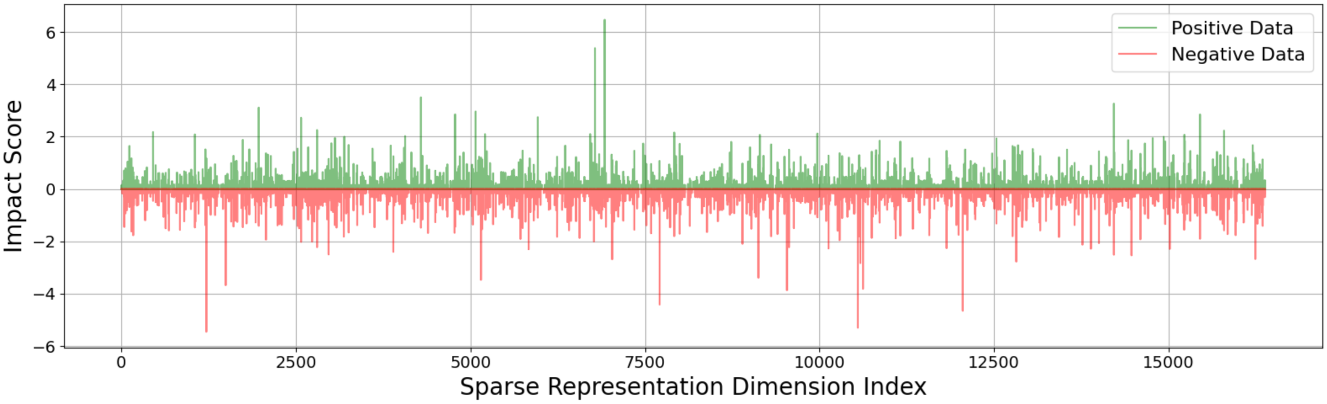

Figure 6: Scores of the positive and negative impacts of each sparse representation dimension in **fairness** domain.

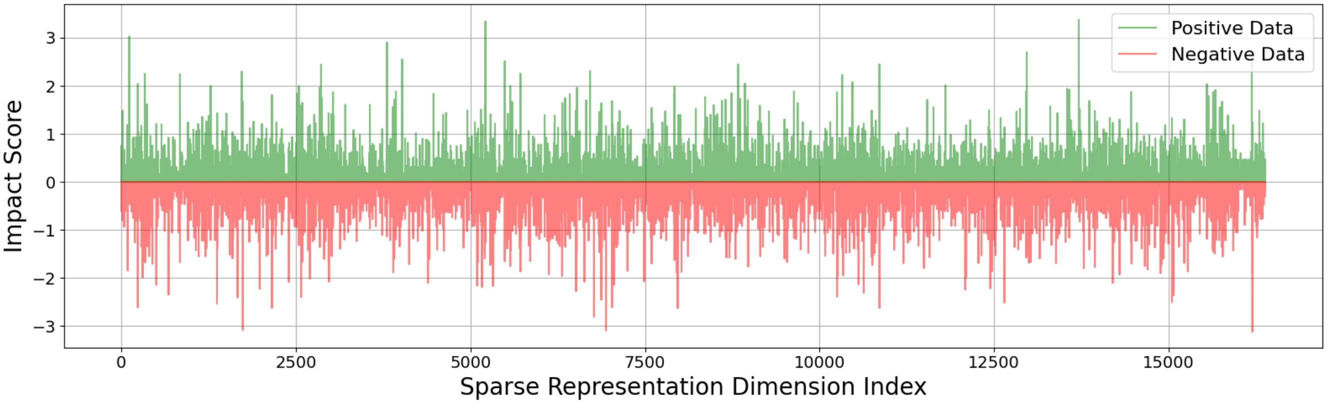

Figure 7: Scores of the positive and negative impacts of each sparse representation dimension in **truthfulness** domain.

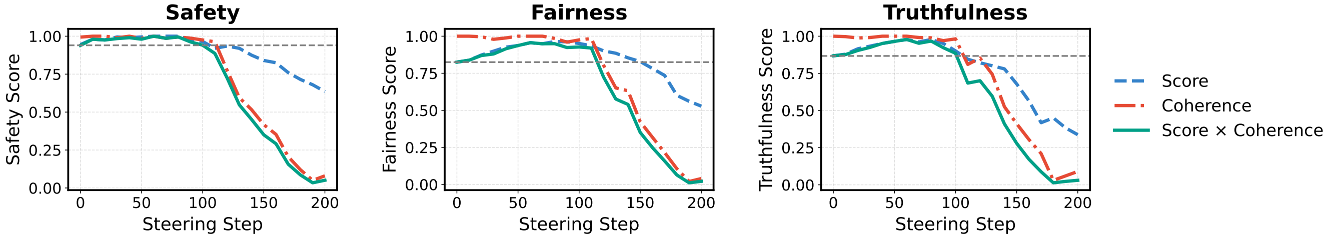

Figure 8: Effect of SRS in three domains with steering scale ranging from 0 to 200 with step=10 on Gemma-2B-it.

property control. However, Coherence exhibits a decline when α becomes too large: while Coherence remains above 0.95 for moderate values ($\alpha \leq 50$), it drops to around 0.90 once ($\alpha \geq 90$), suggesting that strong interventions compromise semantic fidelity. By examining the combined metric Score*Coherence, we identify an optimal trade-off: the product reaches its maximum around $\alpha = 60$, where the model achieves both high property scores and stable coherence. Overall, we set $\alpha = 60$ for all tasks by default.

3. Explanations for Top-K Identified Features

Neuronpedia explanation results on fairness and truthfulness domains are shown in Tab. 8 and Tab. 9, respectively. As shown in Tab. 8, the K_+ features primarily correspond to semantics related to social harmony, positive emotions, and respectful or inclusive expressions, reflecting contexts that promote fairness and equality. In contrast, the K_- features are dominated by concepts associated with negative judgment, discrimination, or hostile descriptions of individuals or groups, which are negatively correlated with fair and unbiased outputs.

As shown in Tab. 9, the sparse features with high positive weights (K_+) tend to capture linguistic patterns related to epistemic uncertainty, logical conditions, and causal reasoning, such as references to “knowledge,” “existence,” and “if-then” structures. These features generally correlate with cautious, truth, seeking language that avoids overstatement. Conversely, the negatively weighted features (K_-) are more aligned with formulaic or technical language, including statistical expressions, programming, related terminology, or document structural markers. These may reflect abstract or rigid content that lacks clear grounding in factual assertions, thereby being less associated with truthful or sincere expression in conversational contexts.

Group	Rank	Index	Explanation of SAE Feature	Weight
K_+	1	6923	elements related to social dynamics and interpersonal relationships	6.46
	2	6784	expressions of happiness and celebration	5.38
	3	4291	conditional or situational phrases in contexts that may imply restrictions or guidelines	3.51
	4	14216	phrases that indicate classifications or types with a focus on personal experiences	3.26
	5	1968	positive descriptors and evaluations of people or things	3.11
	6	5074	features related to medical terminology and health conditions	2.96
	7	4781	references to emotional states and interpersonal relationships	2.85
K_-	1	1218	negative descriptions and issues related to outcomes and performance	-5.45
	2	10549	negative sentiments or harmful concepts in various contexts	-5.30
	3	12051	negative descriptions or reviews of experiences	-4.65
	4	7710	incidents of crime and violence depicted in a societal context	-4.41
	5	9534	concepts related to negative outcomes and their implications	-3.87
	6	10623	topics related to societal judgment and stigma, particularly concerning women and parenting	-3.81
	7	1495	negative descriptors and insults directed toward individuals or groups	-3.67

TABLE 8: Top sparse features identified by SRS for the **fairness** domain on `Gemma-2-2B-it`. Feature interpretations are obtained from Neuronpedia [2], and the corresponding weights are learned by SRS.

Group	Rank	Index	Explanation of SAE Feature	Weight
K_+	1	13713	inquiries about knowledge and uncertainty regarding events or situations	3.37
	2	5215	discussions of legal and social concepts related to guilt and innocence	3.34
	3	114	references to the presence or absence of specific entities or conditions in various contexts	3.02
	4	3805	conditional statements checking for variable existence or conditions	2.90
	5	12968	conjunctions and transition words indicative of causal relationships	2.69
	6	4022	references to events or actions related to conflict, struggle, or disruption in narrative contexts	2.55
	7	16191	digital and numerical data representations	2.52
K_-	1	16200	terms and concepts related to statistical methods and analysis	-3.12
	2	6941	terms related to scientific studies and methodologies	-3.09
	3	1740	references to data processing and analysis methodologies	-3.08
	4	6770	questions and discussions regarding product features and their implications	-2.81
	5	7968	the beginning of sections and titles in structured documents	-2.63
	6	10858	phrases related to safety measures and inventions designed to prevent accidents	-2.62
	7	2157	keywords and identifiers related to programming and software development concepts	-2.61

TABLE 9: Top sparse features identified by SRS for the **truthfulness** domain on `Gemma-2-2B-it`. Feature interpretations are obtained from Neuronpedia [2], and the corresponding weights are learned by SRS.