Instance-Semantic Attribution for Activation Steering in Large Language Models

Anonymous ACL submission

Abstract

Activation Steering guides large language models (LLMs) toward a target behavior by injecting vectors into the residual stream during inference, typically derived from behavioral instances consisting of a prompt paired with positive and negative responses. However, we observe that not all instances contribute equally: while some yield strong and consistent steering signals, others are noisy, unstable, or even misleading, and treating all instances uniformly can degrade overall steering performance. To address this issue, we propose Semantic-Aware Activation Steering (SAAS), a principled framework that performs instance-semantic attribution to assign differentiated importance to instance-level steering vectors based on their relevance to the target behavior. SAAS first assesses instance eligibility by measuring activation separability and directional consistency, and subsequently performs instance selection to retain only stable, behavior-relevant signals. It then uses LLM-based agents to decompose the target behavior into fine-grained semantic sub-behaviors and assess each instance's alignment with them. Finally, the global steering vector is obtained via a weighted aggregation of instance-level vectors, with weights determined by these semantic alignment scores. Experiments across diverse behavior steering tasks demonstrate that SAAS consistently improves steering effectiveness and stability compared to existing activation steering methods. Our code is available at https://anonymous.4open.science/r/SAAS.

1 Introduction

Activation Steering (AS) is a technique for aligning LLMs with target behaviors by injecting steering vectors into the model's activations during inference. It has been applied to tasks such as persona modification (Rimsky et al., 2024; Turner et al., 2023; Cao et al., 2024) and safety enhancement (Lee et al., 2024; Wang et al., 2025; Arditi et al., 2024). Compared to prompting-based approaches, AS is more stable, requires no additional context length, and adds no inference overhead. Compared to fine-tuning, it is simpler to implement and avoids the high computational cost of retraining.

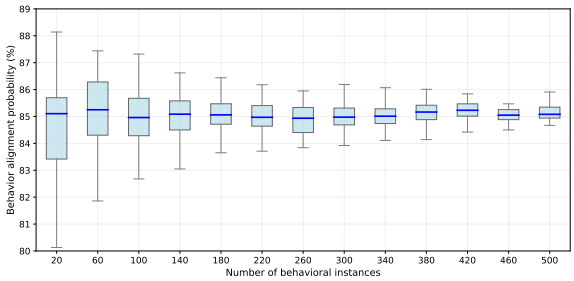

Figure 1: Steering performance on the target behavior (*Conscientiousness*) using steering vectors obtained by averaging randomly sampled subsets of behavioral instances with varying sizes.

The effectiveness of AS hinges on the construction of a high-quality *steering vector* that captures the desired behavioral direction in the model's internal activation space. This vector is typically derived from a collection of *behavioral instances*, where each instance consists of a behavior-related question paired with two answers: a positive answer that exhibits the target behavior, and a negative answer that does not. By computing and analyzing the activation differences between the positive and negative answers, a direction corresponding to the intended behavioral shift can be extracted.

Existing activation steering methods typically rely on two implicit design assumptions: constructing steering vectors from as many behavioral instances as possible, and assigning equal weight to all instances during aggregation. While this strategy promotes stability through averaging, it also enforces a uniform treatment of heterogeneous signals.

Figure 2: An illustrative example of evaluating instance-level semantic attribution through behavior decomposition.

Our empirical analysis reveals two important observations.

First, as shown in Figure 1, although using more instances leads to more stable steering effects, smaller instance sets can occasionally yield stronger alignment, suggesting that simple averaging improves robustness at the cost of peak steering effectiveness.

Second, target behaviors are typically complex and multifaceted, such that different instances may capture different aspects of the same behavior, contributing unequally to the overall behavioral objective. As illustrated in Figure 2, the importance of different instances should be determined by jointly considering these factors.

To address this issue, we propose **Semantic-Aware Activation Steering (SAAS)**, a principled framework that performs instance-level semantic attribution to adaptively modulate the contribution of each steering vector based on its relevance to the target behavior. SAAS first determines instance eligibility by jointly evaluating activation separability and directional consistency, selecting only stable and behaviorally informative signals. It then leverages LLM-based agents to decompose the target behavior into fine-grained semantic sub-behaviors and to assess each instance's alignment with these components. Finally, the global steering vector is obtained by aggregating instance-level vectors through an importance-weighted mechanism informed by these semantic alignment estimates.

Our contributions are as follows:

- We identify a fundamental limitation of existing activation steering methods: uniform treatment of behavioral instances overlooks heterogeneous semantic contributions and the intrinsically multi-faceted structure of target behaviors, limiting steering effectiveness and robustness.
- We propose **Semantic-Aware Activation Steering (SAAS)**, a principled framework that performs instance-level semantic attribution and importance weighting to construct more accurate and stable steering vectors.
- Through extensive experiments across diverse behavior steering tasks, we demonstrate that SAAS consistently improves both steering effectiveness and stability over prior activation steering approaches.

2 Activation Steering

Activation steering modifies internal activations during inference, typically following these steps.

The steering vector $\boldsymbol{v}_l$ at layer l is typically derived from a set of N behavioral instances, where each instance consists of a question q_i paired with two answer tokens: a positive answer y_i^+ that exhibits the target behavior, and a negative answer y_i^- that does not. For simplicity, we omit the layer index l and assume that all operations are performed at the l-th layer.

The most widely adopted approach for extracting the steering vector $\boldsymbol{v}$ is the **Mean Difference (MD)** method (Li et al., 2023; Rimsky et al., 2024; Turner et al., 2023; Arditi et al., 2024). MD computes the average activation difference between paired positive and negative prompts:

$$\boldsymbol{v} = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{a}_i^{\mathrm{p}} - \boldsymbol{a}_i^{\mathrm{n}}\right), \tag{1}$$

where $\boldsymbol{a}_i^{\mathrm{p}} = \boldsymbol{a}(q_i \,||\, y_i^+)$ and $\boldsymbol{a}_i^{\mathrm{n}} = \boldsymbol{a}(q_i \,||\, y_i^-)$. Here, $||$ denotes concatenation, and $\boldsymbol{a}(\cdot)$ denotes the activation of the last token at layer l.

Once extracted, the steering vector is incorporated into the hidden state to steer the generation process. Specifically, at each decoding step during inference, the hidden state corresponding to the current token is updated as follows:

$$\boldsymbol{h}' \leftarrow \boldsymbol{h} + \alpha \cdot \boldsymbol{v}, \tag{2}$$

where α is a scaling hyperparameter that controls the steering strength.

This intervention biases the model's internal activations during inference, thereby increasing the likelihood that its generated output exhibits the desired behavioral trait.

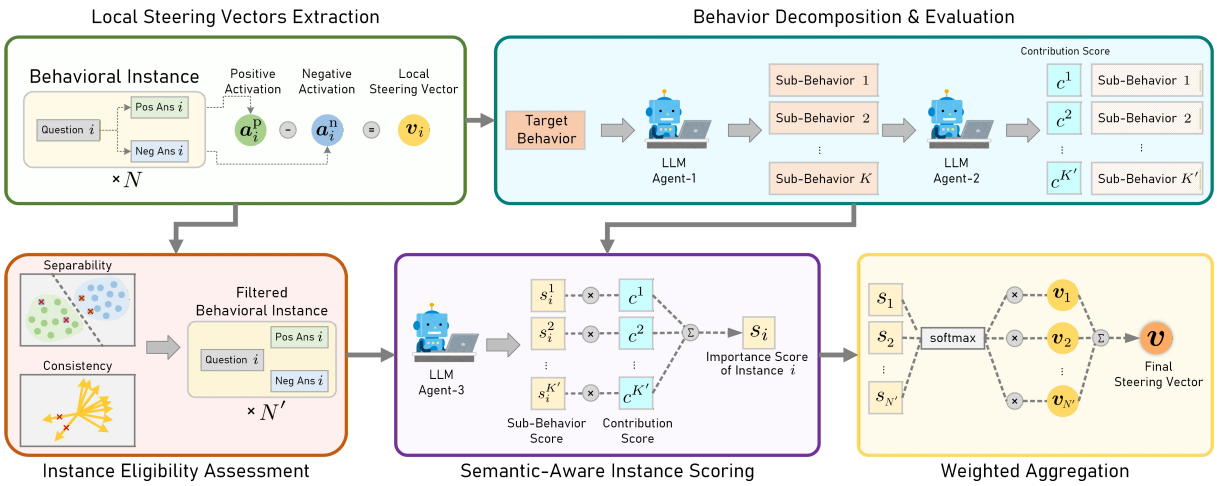

Figure 3: An overview of the proposed framework **SAAS**.

3 Semantic-Aware Activation Steering

We propose **Semantic-Aware Activation Steering (SAAS)**, a framework that constructs a global steering vector by aggregating instance-specific *local* steering vectors using semantic-aware weights that quantify each instance's marginal contribution to inducing the target behavior. Figure 3 provides an overview of the SAAS framework. Implementation details, including the LLM-based agents and their prompts, are available in our codebase.[1]

3.1 Local Steering Vector Extraction

Following the procedure in Section 2, SAAS first extracts the positive and negative activations and computes a local steering vector for each instance:

$$\boldsymbol{v}_i = (\boldsymbol{a}_i^{\mathrm{p}} - \boldsymbol{a}_i^{\mathrm{n}}) \tag{3}$$

This local vector represents the instance-specific activation direction that moves the model's behavior from the negative response toward the positive one for query q_i.

3.2 Instance Eligibility Assessment

The goal of this stage is to determine which behavioral instances are *eligible* to contribute reliable steering signals, while excluding statistically unstable instances.

Inspired by LayerNavigator (Sun et al., 2025), we assess instance eligibility from two complementary perspectives: *activation separability* and *steering vector consistency*.

[1] https://anonymous.4open.science/r/SAAS

Activation Separability. An instance is considered eligible only if it provides a reliable contrast between its positive and negative responses. To quantify this property, we evaluate the separability of their activations using a Fisher linear discriminant. Specifically, we compute a discriminant direction that maximally separates positive and negative activations across all instances. Instances whose positive or negative activations project within half of the mean separation distance along the mean difference direction are deemed insufficiently separable and are excluded from further consideration.

The discriminant direction is given by:

$$\boldsymbol{w}_{\text{sep}} = \boldsymbol{S}_w^{-1}(\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-), \tag{4}$$

where $\boldsymbol{S}_w$ denotes the within-class scatter matrix, and $\boldsymbol{\mu}^+$ and $\boldsymbol{\mu}^-$ are the mean activations of positive and negative activations, respectively.

Steering Vector Consistency. In addition to separability, eligible instances are required to exhibit directional consistency with the dominant steering trend. To this end, we compute the mean vector of all local steering vectors $\boldsymbol{v}_i$, and remove instances whose local steering vectors form an obtuse angle (greater than $90°$) with the global direction.

After this assessment, N' eligible instances are retained for subsequent semantic attribution and weighting.

3.3 Behavior Decomposition & Instance Scoring

To assign fine-grained semantic relevance scores to the remaining N' instances in an efficient and interpretable manner, SAAS employs LLM-based

agents as semantic evaluators. These agents are suited for capturing high-level behavioral semantics and can provide explicit, human-interpretable rationales for relevance assessment.

Agent-1: Behavior Decomposition. SAAS begins by introducing **Agent-1**, which decomposes the target behavior into a set of fine-grained sub-behaviors. This design is motivated by the observation that instance-level relevance is often difficult to assess with respect to a complex and multifaceted behavior. By decomposing the target behavior, relevance evaluation can be conducted at a more interpretable and semantically discriminative granularity.

Concretely, Agent-1 is provided with the target behavior description together with five randomly sampled training instances as illustrative examples. Based on this input, Agent-1 is instructed to generate $K \in [5, 10]$ sub-behaviors that satisfy the following criteria: (i) each sub-behavior is concrete and behaviorally specific rather than abstract; (ii) the sub-behaviors are mutually non-overlapping; and (iii) they are not tied to any individual training instance, but instead capture generalizable aspects of the target behavior.

Agent-2: Sub-behavior Evaluation. Sub-behaviors obtained via naive decomposition can differ substantially in their relevance to the target behavior. To address this issue, SAAS introduces a second agent, **Agent-2**, which evaluates the contribution of each candidate sub-behavior.

Notably, Agent-2 does not observe any training instances. Instead, it is provided solely with the target behavior description and the set of candidate sub-behaviors. From sociological and behavioral perspectives, Agent-2 assigns each sub-behavior a contribution score $c^k \in [0, 10]$, reflecting the extent to which the sub-behavior characterizes or supports the target behavior. Sub-behaviors with scores below 2 are discarded, while the remaining K'sub-behaviors are retained for subsequent instance-level relevance scoring.

Together, this two-stage agent design serves complementary roles: Agent-1 expands the semantic space by decomposing complex behaviors into interpretable components, while Agent-2 performs a conservative pruning step that filters out weak or spurious sub-behaviors without relying on instance-specific evidence.

Following the previous stages, SAAS retains N' filtered instances along with K' validated sub-behaviors. We then introduce a third agent, **Agent-3**, to perform semantic-aware instance scoring.

Agent-3: Semantic-Aware Instance Scoring For each instance i, Agent-3 evaluates its semantic relevance with respect to each sub-behavior k and assigns a relevance score $s_i^k \in [0, 10]$. Agent-3 is provided with both the instance content and the description of the corresponding sub-behavior, and is instructed to assess the extent to which the instance exhibits that sub-behavior.

To obtain a single importance score for each instance, the relevance scores across all sub-behaviors are aggregated using their contribution weights. Specifically, the final importance score for instance i is computed as:

$$s_i = \sum_{k=1}^{K'} c^k \times s_i^k. \tag{5}$$

The resulting score s_i reflects both the semantic relevance of instance i to individual sub-behaviors and the relative importance of those sub-behaviors to the target behavior.

3.4 Weighted Aggregation

Finally, the instance-level importance scores are converted into normalized weights, which are then used to aggregate the local steering vectors:

$$w_i = \frac{\exp(s_i/t)}{\sum_{j=1}^{N'} \exp(s_j/t)}, \tag{6}$$

$$\boldsymbol{v} = \sum_{i=1}^{N'} w_i \times \boldsymbol{v}_i, \tag{7}$$

where $t > 0$ is a temperature parameter that controls the sharpness of the distribution. This weighted aggregation prioritizes instances that are both semantically aligned and statistically reliable, effectively reducing the influence of noisy or weakly relevant examples and producing a steering vector that robustly represents the target behavior.

4 Experiments

4.1 Experimental Settings

4.1.1 Baselines

We use Llama-3-8B-Instruct (Dubey et al., 2024) as the base models and compare our method against the following baselines:

- **CAA** (Rimsky et al., 2024): Computes steering vectors using the Mean Difference approach (see Section 2), capturing the average directional shift between positive and negative behavior instances.

- **PCA-center** (Lee et al., 2024; Adila et al., 2024): Extracts the first principal component from all positive and negative activations after centering, identifying the direction of maximal variance across activations.

- **PCA-diff** (Zou et al., 2023; Ball et al., 2024): Derives the first principal component from local steering vectors, representing the dominant direction among these instance-level differences.

- **BiPO** (Cao et al., 2024): Optimizes a steering vector to directly maximize the likelihood of target behavior outputs while minimizing that of opposing behaviors.

We additionally evaluate the effectiveness of weighted aggregation by comparing SAAS with two alternative heuristic weighting strategies:

- **logit-diff**: Assigns weights to instances based on the difference in logits between the model's preferred positive and negative tokens, reflecting the model's relative preference.

- **conf**: Uses the model's softmax confidence on the positive token as the instance weight, capturing the absolute certainty of the model in each instance.

4.1.2 Behaviors and Datasets

AI Persona. We evaluate SAAS on persona steering using Anthropic's Persona Dataset (Perez et al., 2023), which provides 1,000 behavior-related questions per persona. Each question is paired with a binary (`Yes`/`No`) label indicating whether the model's response aligns with the target persona. Our evaluation focuses on the *OCEAN* personality traits—*Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*—the Big Five dimensions widely adopted in psychology and increasingly used to assess persona control in LLMs (Jiang et al., 2023).

Safety Enhancement. We further test SAAS on two safety-critical behaviors from Rimsky et al. (2024): *Refusal*, which evaluates the model's ability to appropriately decline harmful or disallowed requests, and *Hallucination*, which measures robustness against generating factually incorrect information in response to misleading prompts. Each task consists of 1,000 instances, with each instance presenting a prompt and two candidate responses—one aligned with the target behavior and one misaligned. Both datasets involve open-ended questions and evaluate behavioral alignment at the response level.

4.1.3 Experimental Setup

For each behavior, we randomly split the dataset into training, validation, and test sets (7:2:1). Steering vectors are derived from the training set, and the steering layer l and softmax temperature t are selected based on validation performance (see Section 4.5). All experiments are repeated over five random splits, and test set results are reported as the mean.

4.1.4 Evaluation Metrics

For AI persona tasks, following (Rimsky et al., 2024), we measure the model's alignment with the target persona using the average token-level **probability** assigned to behavior-consistent responses.

For safety-enhancement tasks, following (Cao et al., 2024), we employ GPT-4 as an automatic evaluator to compute a **behavioral score**. GPT-4 rates each response on a 1–4 scale according to behavior-specific criteria, and the final score is averaged across all prompts to quantify behavioral alignment.

4.2 AI Persona

Table 1 reports behavior alignment probabilities on the *OCEAN* personality traits under both negative ($\alpha = -1.0$) and positive ($\alpha = +1.0$) steering, corresponding to suppressing and enhancing the target trait, respectively.

Across all five personality dimensions, **SAAS** consistently achieves the strongest overall steering performance. In the positive steering setting ($\alpha = +1.0$), SAAS attains the best alignment scores for all traits, indicating its superior ability to enhance target persona characteristics in a reliable and uniform manner. The improvements are particularly pronounced for *Extraversion* and *Neuroticism*, where SAAS yields clear gains over the strongest competing methods.

Under negative steering ($\alpha = -1.0$), SAAS also produces the lowest alignment probabilities across all traits, demonstrating effective suppression of

Method	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
	$-1.0\downarrow$	$+1.0\uparrow$	$-1.0\downarrow$	$+1.0\uparrow$	$-1.0\downarrow$	$+1.0\uparrow$	$-1.0\downarrow$	$+1.0\uparrow$	$-1.0\downarrow$	$+1.0\uparrow$
Base	93.05		79.82		60.82		90.78		66.41	
CAA	85.41	93.35	71.07	84.30	60.36	60.44	75.25	95.23	61.66	69.47
PCA-center	91.05	94.08	73.31	82.86	59.64	60.49	89.72	91.06	61.62	68.25
PCA-diff	91.51	94.00	72.18	82.90	59.56	60.59	89.70	91.26	61.52	68.41
BiPO	86.07	94.11	70.03	84.57	58.47	60.93	76.72	94.53	60.23	70.48
logit-diff	91.17	93.82	72.65	83.07	58.65	60.66	87.54	93.97	61.10	67.95
conf	90.83	93.56	73.24	82.21	58.23	60.32	82.07	94.13	61.06	67.42
SAAS	**83.12**	**95.65**	**68.45**	**86.32**	**56.21**	**64.21**	**72.30**	**97.31**	**56.11**	**74.41**

Table 1: Behavior Alignment Probability (%) of OCEAN personalities under different steering strengths $\alpha \in \{-1.0, +1.0\}$. The best results are shown in **bold**, and the second-best are underlined.

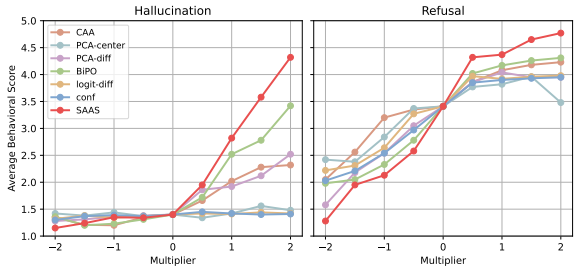

Figure 4: The comparison results on steering *Refusal* and *Hallucination* tasks.

undesired persona attributes. Importantly, this behavior is consistent across dimensions, suggesting that SAAS enables precise bidirectional control rather than relying on asymmetric amplification effects.

In contrast, existing baselines exhibit less reliable performance and, in certain settings—such as *Extraversion*—struggle to produce an effective steering vector. These results indicate that uniform or heuristic instance aggregation strategies are insufficient for robust persona steering, highlighting the advantages of SAAS's instance-level semantic attribution and adaptive aggregation.

4.3 Refusal and Hallucination

Figure 4 compares the steering performance of different methods on the *Refusal* and *Hallucination* tasks under varying steering strengths. Both tasks evaluate safety-critical behaviors, where effective control requires not only performance gains but also stability across steering magnitudes.

Overall, SAAS consistently outperforms all baseline methods on both tasks. As the steering strength α increases, SAAS exhibits smooth and monotonic improvements in average behavioral scores, indicating stable and predictable steering behavior. In contrast, baseline methods often plateau at moderate steering strengths or display noticeable fluctuations, suggesting limited robustness to stronger interventions.

The advantage of SAAS is particularly evident at higher values of α, where uncontrolled steering may amplify noise or induce undesired side effects. SAAS maintains steady performance gains without abrupt degradation, demonstrating its ability to reliably enhance safety behaviors such as refusal and hallucination mitigation.

4.4 Ablation Study

We conduct an ablation study on the *OCEAN* personality traits to systematically examine the contribution of each component in SAAS.

Effect of Instance Eligibility Assessment. We first examine the impact of instance eligibility assessment. As shown in Table 2, approximately 1%–9% of instances are deemed ineligible across different traits. Reintroducing these ineligible instances leads to an average performance drop of about 1%, indicating that eligibility assessment effectively removes noisy or harmful instances and is important for stable steering.

Effect of Instance-Level Semantic Weighting. We next evaluate the contribution of LLM-based instance weighting by replacing semantic-aware weighting with uniform averaging over eligible instances. As shown in Table 2, performance decreases substantially, confirming that instance-level semantic attribution plays a key role beyond eligibility assessment alone.

Setting	O	C	E	A	N
Eligible instances (%)	98.5	96.4	91.7	92.5	93.1
No eligibility assessment (Δ%)	-0.82	-0.76	-1.32	1.54	-1.05
No agents (uniform averaging) (Δ%)	-1.62	-2.95	-3.20	-1.63	-1.07
No Agent-2 (no sub-behavior evaluation) (Δ%)	-0.78	-1.08	-0.62	-0.54	-1.13
Agent = DeepSeek-V3.2 (Δ%)	-0.23	+0.54	+0.22	-0.16	-0.10
Agent = Gemini-2.5 (Δ%)	+0.24	+0.12	-0.08	-0.12	+0.25

Table 2: Ablation results of SAAS on the OCEAN personality traits. Δ% denotes relative performance change compared to the full SAAS model.

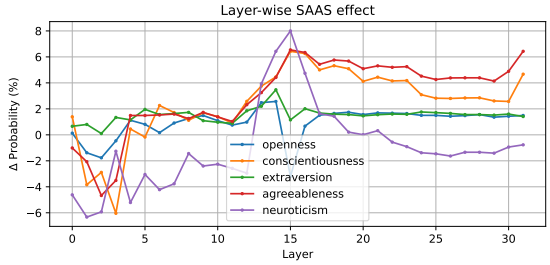

Figure 5: Effect of SAAS at different layers ($\alpha = 1.0$).

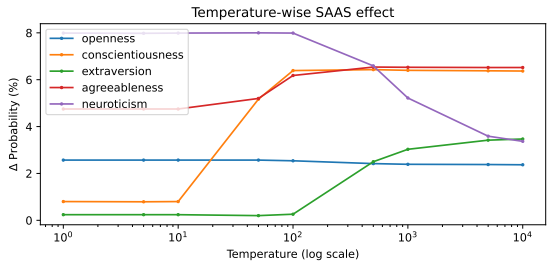

Figure 6: Effect of SAAS with different temperatures t ($\alpha = 1.0$).

Effect of Sub-behavior Evaluation (Agent-2). To assess the role of sub-behavior evaluation, we disable Agent-2 and assign equal weights to all sub-behaviors produced by Agent-1. This results in a noticeable performance degradation, demonstrating the necessity of selecting semantically relevant sub-behaviors.

Robustness to Agent Model Choice. Finally, we evaluate robustness to different agent LLMs by replacing the default agent with DeepSeek-V3.2 (Liu et al., 2025) and Gemini-2.5 (Comanici et al., 2025). Performance differences are marginal, suggesting that SAAS is insensitive to the specific choice of modern LLMs for instance scoring and sub-behavior analysis.

4.5 Hyperparameters Study

We conduct hyperparameter experiments on the validation set and observe the following patterns.

As shown in Figure 5, unlike other baselines such as CAA, which exhibit relatively smooth and consistent performance trends across layers, SAAS shows irregular fluctuations in the earlier layers. This pattern can be explained by the eligibility assessment stage: at shallow layers, target behavior information is weakly encoded, causing most instances to be deemed ineligible and removed. In practice, activation steering methods tend to be ineffective at early layers, and SAAS makes this limitation more apparent through explicit eligibility assessment. It is generally acknowledged that activation steering tends to perform best at middle-to-late layers, and SAAS is consistent with this observation, achieving its optimal performance at layers 14–15.

We additionally evaluate the effect of the temperature parameter, as shown in Figure 6. The results indicate that certain behaviors benefit from a higher temperature—corresponding to smoother aggregation—such as *Conscientiousness*, whereas others favor sharper aggregation with lower temperatures, such as *Neuroticism*. We hypothesize that this discrepancy primarily stems from intrinsic differences in the training data. When the data quality is relatively high, smoother aggregation is more appropriate; conversely, in noisier or lower-quality settings, SAAS benefits from sharper weighting and demonstrates stronger effectiveness.

4.6 Effect of Steering Strength and Layers

We investigate how the number of steering layers L and the steering strength α jointly affect SAAS performance (Figure 7). Specifically, we select the top-L layers that perform best on the validation set when performing single-layer steering.

Both L and α show non-monotonic patterns: increasing either improves behavior alignment up to a point, but further increases cause performance degradation, indicating oversteering risks. Interestingly, L and α exhibit a compensatory relationship:

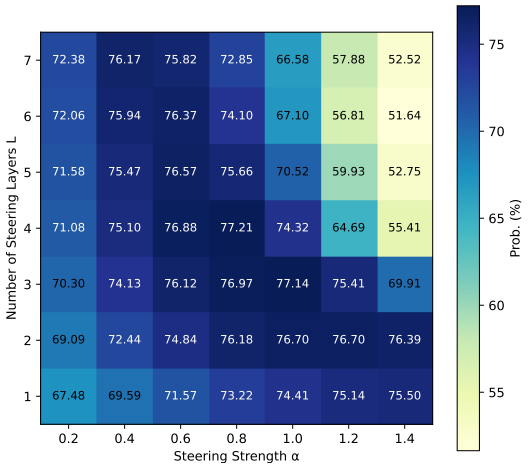

Figure 7: Behavioral control performance on the *Neuroticism* persona under varying steering strength α and number of steering layers L.

Behavior	+1.0	-1.0
Openness	66.4	66.3
Conscientiousness	66.1	66.4
Extraversion	66.0	66.2
Agreeableness	66.1	66.5
Neuroticism	66.3	66.0
Refusal	66.8	66.1
Hallucination	65.9	66.7

Table 3: MMLU accuracy (%) under different steering multipliers for each behavioral persona. The baseline accuracy is 66.4%.

strong steering at few layers (e.g., $L = 2$, $\alpha = 0.8$) achieves similar results to weaker steering across more layers (e.g., $L = 7$, $\alpha = 0.4$). This complementarity enables more flexible and stable control strategies.

4.7 Impact on General Utility

To assess the impact of SAAS on the model's general knowledge and problem-solving capabilities, we evaluate it on the MMLU benchmark (Hendrycks et al., 2021), which includes multiple-choice questions from 57 diverse professional subjects. For each subject, we randomly sample 20 questions and measure accuracy under different steering directions.

As shown in Table 3, SAAS has a negligible impact on overall performance. This indicates that SAAS preserves the model's utility across tasks, thanks to its instance-weighting mechanism, which selectively amplifies behavior-relevant signals while suppressing unrelated noise.

4.8 Related Work

Behavior Steering in Large Language Models Behavior steering refers to the task of modulating LLM outputs to express or suppress specific behaviors. Closely related terms such as *control* and *alignment* are often used interchangeably in the literature, as they largely share the same objective. Prior work has explored diverse objectives, including promoting truthfulness (Li et al., 2023; Qiu et al., 2024; Wang et al., 2024a,b), aligning with human values (Zou et al., 2023; Cheng et al., 2024), refusing unsafe requests (Lee et al., 2024; Arditi et al., 2024; Yousefpour et al., 2025), role-playing (Louie et al., 2024), and simulating specific persona traits (van der Weij et al., 2024; Rimsky et al., 2024; Jiang et al., 2023; Choi and Li, 2024).

Activation Steering Activation steering has gained traction due to its efficiency and interpretability. Early methods such as Activation Addition (ActAdd)(Turner et al., 2023) derive steering vectors from single contrastive examples. More scalable variants like Inference-Time Intervention (ITI)(Li et al., 2023) and Contrastive Activation Addition (CAA) (Rimsky et al., 2024) extract the Mean Difference (MD) vector across datasets of positive and negative responses. Conditional Activation Steering (CAST) (Lee et al., 2024) uses PCA-based vectors to implement selective refusals.

While recent studies (Tigges et al., 2023; Zou et al., 2023) affirm the empirical strength of AS, they also highlight its limitations. In particular, Tan et al. (2024) identifies its limitations, showing that existing approaches are sensitive to instance quality and may fail in the presence of weak or noisy examples. These observations motivate our work, which rethinks vector construction through instance-level weighting.

5 Conclusion

We propose Semantic-Aware Activation Steering (SAAS), an instance-semantic attribution framework for constructing reliable steering vectors in large language models. Experiments across diverse behavior steering tasks demonstrate that SAAS consistently improves steering effectiveness and stability over strong baselines. Ablation studies confirm the contribution of each component, while additional analyses show that SAAS preserves general model capabilities and remains robust across different steering layers and strengths.

6 Limitations

Our method focuses on identifying and assigning higher weights to instances with strong steering potential. While these instances are drawn from widely used behavioral datasets, their quality is inherently limited, as such datasets are primarily designed to evaluate whether LLMs exhibit certain behaviors, rather than to induce or steer them. As a result, some selected instances may still contain noise, ambiguity, or weak behavioral signals. We view the development of higher-quality, steering-oriented instances as an important future direction, which could address this limitation at the data source and further enhance the effectiveness of SAAS.

7 Ethical Considerations

SAAS is a general activation-based steering framework and, like other behavior steering techniques, may be misused to shift large language models toward unsafe, biased, or otherwise undesirable behaviors if applied without appropriate safeguards. In particular, the ability to selectively amplify specific behavioral signals could be exploited to reinforce harmful tendencies or bypass intended model constraints.

We emphasize that our work is intended to support controlled, transparent, and research-oriented behavior steering, such as improving safety, reliability, and interpretability of LLMs. Responsible use of SAAS should be accompanied by proper dataset curation, clear behavioral definitions, and downstream safety evaluations. We hope that this work encourages further research on developing robust safeguards and governance mechanisms for activation-level interventions in large language models.

References

Dyah Adila, Shuai Zhang, Boran Han, and Yuyang Wang. 2024. Discovering bias in latent space: An unsupervised debiasing approach. *arXiv preprint arXiv:2406.03631*.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083.

Sarah Ball, Frauke Kreuter, and Nina Panickssery. 2024. Understanding jailbreak success: A study of latent space dynamics in large language models. *arXiv preprint arXiv:2406.09289*.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. *Advances in Neural Information Processing Systems*, 37:49519–49551.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2024. Black-box prompt optimization: Aligning large language models without model training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3201–3219.

Hyeong Kyu Choi and Yixuan Li. 2024. Picle: Eliciting diverse behaviors from large language models with persona in-context learning. In *International Conference on Machine Learning*, pages 8722–8739. PMLR.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643.

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. 2024. Programming refusal with conditional activation steering. *arXiv preprint arXiv:2409.05907*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10570–10603.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Maria Ponti, and Shay Cohen. 2024. Spectral editing of activations for large language model alignment. *Advances in Neural Information Processing Systems*, 37:56958–56987.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Hao Sun, Huailiang Peng, Qiong Dai, Xu Bai, and Yanan Cao. 2025. Layernavigator: Finding promising intervention layers for efficient activation steering in large language models. *Advances in Neural Information Processing Systems*.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. Analysing the generalisation and reliability of steering vectors. *Advances in Neural Information Processing Systems*, 37:139179–139212.

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

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

Teun van der Weij, Massimo Poesio, and Nandi Schoots. 2024. Extending activation steering to broad skills and multiple behaviours. *arXiv preprint arXiv:2403.05767*.

Tianlong Wang, Xianfeng Jiao, Yifan He, Zhongzhi Chen, Yinghao Zhu, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. 2024a. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. *arXiv preprint arXiv:2406.00034*.

Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. 2025. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the ACM on Web Conference 2025*, pages 2562–2578.

Weixuan Wang, Jingyuan Yang, and Wei Peng. 2024b. Semantics-adaptive activation intervention for llms via dynamic steering vectors. *arXiv preprint arXiv:2410.12299*.

Ashkan Yousefpour, Taeheon Kim, Ryan Sungmo Kwon, Seungbeen Lee, Wonje Jeung, Seungju Han, Alvin Wan, Harrison Ngan, Youngjae Yu, and Jonghyun Choi. 2025. Representation bending for large language model safety. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24073–24098.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.