
One SPACE to Rule Them All: Jointly Mitigating Factuality and Faithfulness Hallucinations in LLMs

Pengbo Wang[†] Chaozhuo Li^{†*} Chenxu Wang[‡] Liwen Zheng[†] Litian Zhang[†] Xi Zhang[†]

[†]Beijing University of Posts and Telecommunications

[‡]Shihezi University

{1696280515, lichaozhuo, zhenglw, litianzhang, zhangx}@bupt.edu.cn
axsdsama@gmail.com

Abstract

LLMs have demonstrated unprecedented capabilities in natural language processing, yet their practical deployment remains hindered by persistent factuality and faithfulness hallucinations. While existing methods address these hallucination types independently, they inadvertently induce performance trade-offs, as interventions targeting one type often exacerbate the other. Through empirical and theoretical analysis of activation space dynamics in LLMs, we reveal that these hallucination categories share overlapping subspaces within neural representations, presenting an opportunity for concurrent mitigation. To harness this insight, we propose SPACE, a unified framework that jointly enhances factuality and faithfulness by editing shared activation subspaces. SPACE establishes a geometric foundation for shared subspace existence through dual-task feature modeling, then identifies and edits these subspaces via a hybrid probe strategy combining spectral clustering and attention head saliency scoring. Experimental results across multiple benchmark datasets demonstrate the superiority of our approach.

1 Introduction

LLMs have fundamentally transformed the landscape of natural language processing [1, 2]. However, a fundamental challenge remains unaddressed: the persistent manifestation of hallucinations, defined as instances where these models generate text containing factual incongruities or deviations from contextual coherence [3]. This phenomenon becomes particularly consequential when considering application scenarios demanding rigorous epistemological integrity.

Hallucinations can be classified into two primary types: factuality hallucinations and faithfulness hallucinations [3, 4, 5, 6]. Factuality hallucinations refer to contradictions against verifiable facts, whereas faithfulness hallucinations denote deviations from user intent or exhibit internal inconsistencies. As shown in Figure 1(a), claiming “Sydney is the capital of Australia” is a factuality hallucination, while answering “The cheetah runs fastest” to “Who runs faster, the turtle or the rabbit?” is a faithfulness hallucination, since it is correct but unfaithful to the prompt. Both categories significantly compromise the reliability of LLMs, as even isolated instances erode user trust in generated content.

Existing works generally address hallucination categories in isolation, focusing on optimizing either factuality [7, 8, 9] or faithfulness [10, 11, 12] through distinct methodological frameworks. However, our study reveals a critical oversight: interventions targeting one type of hallucination often exacerbate others. As illustrated in Figure 1(b), experiments are conducted on two benchmark datasets: TruthfulQA [13] for factuality evaluation and the PDTB [14] corpus, with DISQ [15] employed as the

*Corresponding author: Chaozhuo Li

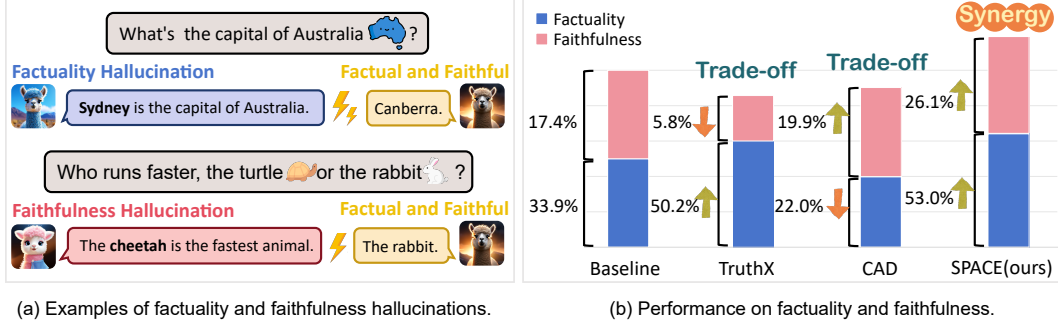


Figure 1: The illustrations of two hallucination types and the performance trade-off.

metric for faithfulness. Two prominent models—TruthX (designed for factual accuracy optimization) and CAD (developed to enhance faithfulness)—are selected as comparative baselines. Compared to the base model (Llama-2-7b), TruthX significantly improves factuality metrics but shows performance degradation when evaluated on the PDTB corpus. Conversely, CAD improves faithfulness on PDTB while sacrificing factuality. These results underscore a clear trade-off between factuality and faithfulness in existing works.

Focusing on the tradeoff between different types of hallucinations, this paper proposes to simultaneously enhance performance for both categories through a unified operation – achieving “killing two birds with one stone”. Building on prior research [16, 17], the activation space is adopted as the technical substrate. Formally defined as the comprehensive set of intermediate representations generated by neural networks, the activation space provides crucial insights into the cognitive mechanisms of LLMs [18] and reveals intrinsic correlations with hallucination phenomena [19, 20]. Hallucination mitigation can be systematically interpreted as targeted modifications within particular subspaces of the activation space [21]. Intuitively, the subspaces associated with different hallucination types appear independent due to their distinct definitions [22, 23]. Editing or expanding the subspace related to one type of hallucination may inadvertently distort or compress the subspace of another, thereby creating performance tradeoffs. However, due to the shared parameter architecture of LLMs [24] and interconnected nature of different hallucination types, their subspaces may exist overlapped region. If such shared subspace could be targeted for editing, both hallucination types could be concurrently optimized, resulting in mutual performance enhancement rather than compromise.

While promising, modifying shared activation subspaces to simultaneously improve factuality and faithfulness in LLMs remains challenging. First, the existence of such a shared subspace is often intuitively assumed, yet providing a rigorous theoretical foundation for its validity remains non-trivial. Second, identifying this subspace is inherently difficult due to the high-dimensional activation space of LLMs, which complicates precise localization. Third, even if such a subspace were identifiable, developing an effective editing methodology that modifies the targeted space without compromising model performance introduces additional difficulty.

To address the aforementioned challenges, we propose SPACE (Spatial Processing for Activated Combined Embeddings), a novel framework that steers LLMs toward the shared activation space of factuality and faithfulness. We rigorously formalize the intersection of activation patterns in Section 2.2 through systematic extraction and geometric modeling of dual-task activation features, thereby establishing a mathematical foundation for subspace existence. A hybrid probe strategy is further proposed to combine the spectral clustering with attention head saliency scoring, enabling precise identification of critical subspace-contributing heads. By constructing adjustment vectors derived from the extracted features, SPACE ensures these edits align with both objectives (factuality and faithfulness) without biasing either. These vectors are then applied to the identified attention heads, enabling precise modifications that preserve overall model performance. Experimental results over SOTA baselines demonstrate the superiority of our proposal.

In summary, our contributions are as follows:

- We theoretically reveal a trade-off between mitigating factuality and faithfulness hallucinations in LLMs, which stems from divergent activation subspaces optimized for each task during training.

- To resolve this trade-off, we propose SPACE, a novel model that identifies and edits the shared activation subspace where factuality and faithfulness intersect, resulting in the “two birds with one stone” effect.
- Extensive experiments on standard benchmarks demonstrate SPACE’s superiority over SOTA baselines, validating its effectiveness in jointly improving LLM factuality and faithfulness.

2 Preliminary Analysis

2.1 Empirical Analysis of Hallucination-Related Activation Patterns

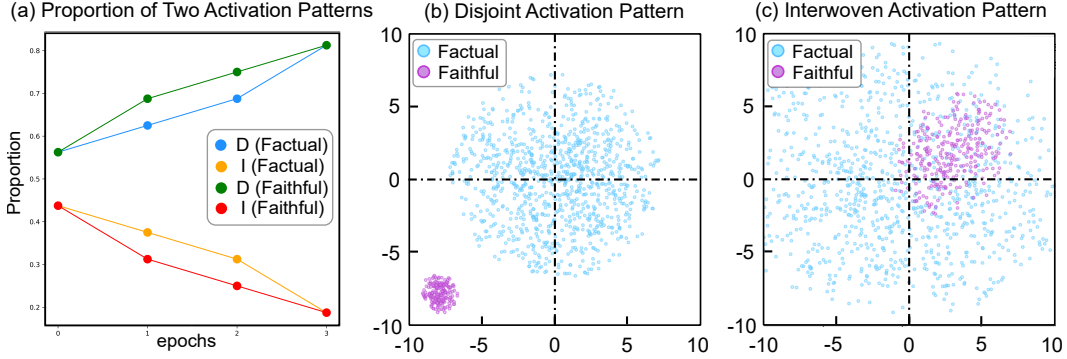


Figure 2: Activation Distributions of Factual and Faithful Tasks. Subfigure (a) shows the proportion of disjoint and interwoven activation pattern after training. Subfigure (b) and (c) show two cases of disjoint activation pattern and interwoven activation pattern respectfully.

Our research commences with a systematic examination of activation dynamics in LLMs concerning their factuality and faithfulness. The study employs Llama3.2-1B as the target model and utilizes TruthfulQA [13] for factual enhancement and PDTB [15] for faithfulness improvement through instruction tuning. In each training epoch, the activations from each layer of the target LLM are collected and visualized via PCA. As illustrated in Figure 2 (b) and (c), we identify two distinct activation patterns: the disjoint activation pattern, which exhibits clear spatial separation between neural activations related to factual and faithful processing, and the interwoven activation pattern, where these two types of activations show significant overlap. During training on the factual dataset, the proportion of disjoint activation patterns (D(Factual)) exhibits a consistent increase, whereas interwoven patterns (I(Factual)) demonstrate a decline. This divergence suggests a progressive separation between the two activation subspaces, indicative of their growing relative distance.

The results imply that activation subspaces corresponding to distinct task types are localized within the model’s parameter space. However, introducing additional tasks during training can perturb these subspaces, potentially destabilizing pre-existing representations. A promising direction for mitigating this interference involves identifying and isolating these subspaces, then guiding the LLM to learn features within their intersectional region. Such an approach could enable the model to harmonize and retain the desired attributes of both tasks in its generative outputs, thereby ameliorating the trade-off.

2.2 Theoretical Guarantee for the Existence of the Overlapped Subspace

The key to realizing the conjecture formulated in the previous section is to ensure the existence of an intersection between the activation subspaces associated with factuality and faithfulness. We provide a theoretical proof that such an overlapped subspace necessarily exists in LLMs.

Let $\mathcal{V} \subseteq \mathbb{R}^d$ be a complete Hilbert space equipped with the standard inner product $\langle \cdot, \cdot \rangle$. Define the factuality subspace $\mathcal{F}_{\text{fact}} \subseteq \mathcal{V}$ and the faithfulness subspace $\mathcal{F}_{\text{faith}} \subseteq \mathcal{V}$ as two nonempty, closed, and convex subsets of \mathcal{V} . Additionally, suppose that these sets satisfy the following convex interpolation property: for any $\mathbf{v} \in \mathcal{F}_{\text{fact}}$ and $\mathbf{u} \in \mathcal{F}_{\text{faith}}$, and for any $\lambda \in [0, 1]$, the convex combination

$$\mathbf{w} = \lambda \mathbf{v} + (1 - \lambda) \mathbf{u}$$

also belongs to $\mathcal{F}_{\text{fact}} \cap \mathcal{F}_{\text{faith}}$. This assumption reflects the preservation of both factuality and faithfulness properties under linear interpolation.

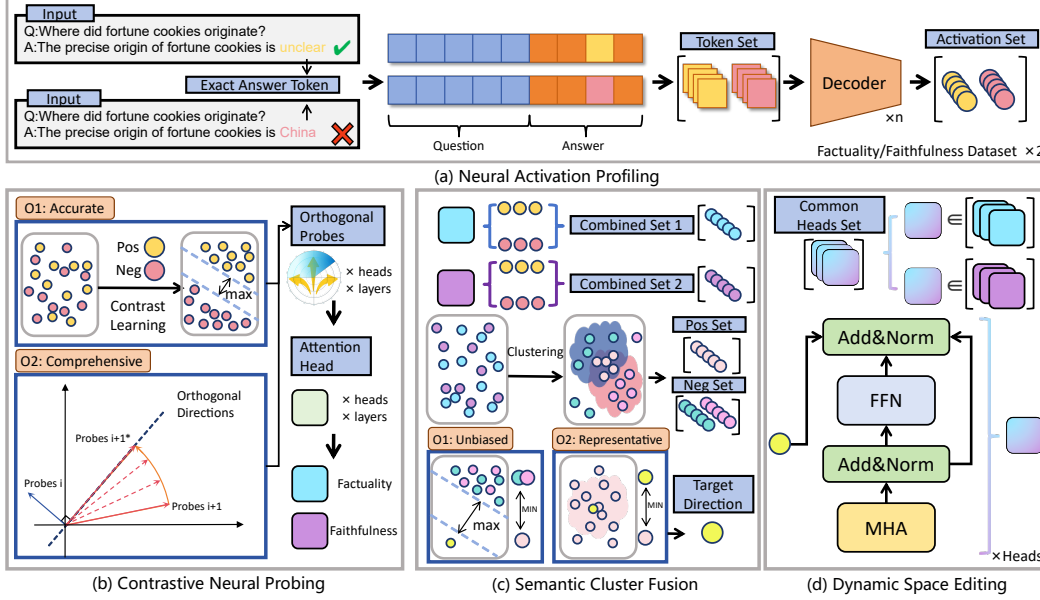


Figure 3: Framework of the proposed SPACE model.

To establish the existence of an intersection between $\mathcal{F}_{\text{fact}}$ and $\mathcal{F}_{\text{faith}}$, suppose, for contradiction, that $\mathcal{F}_{\text{fact}} \cap \mathcal{F}_{\text{faith}} = \emptyset$. By the Hahn–Banach Separation Theorem [25], there exists a nonzero vector $\mathbf{a} \in \mathcal{V}$ and a constant $\gamma \in \mathbb{R}$ such that

$$\langle \mathbf{a}, \mathbf{v} \rangle \leq \gamma < \langle \mathbf{a}, \mathbf{u} \rangle, \quad \forall \mathbf{v} \in \mathcal{F}_{\text{fact}}, \quad \forall \mathbf{u} \in \mathcal{F}_{\text{faith}}.$$

This implies that every point in $\mathcal{F}_{\text{fact}}$ lies entirely in one half-space defined by \mathbf{a} , while every point in $\mathcal{F}_{\text{faith}}$ lies strictly in the other half-space.

Now, consider an arbitrary $\mathbf{v} \in \mathcal{F}_{\text{fact}}$ and $\mathbf{u} \in \mathcal{F}_{\text{faith}}$. By the convex interpolation assumption, for any $\lambda \in (0, 1)$, the point

$$\mathbf{w} = \lambda \mathbf{v} + (1 - \lambda) \mathbf{u}$$

must belong to $\mathcal{F}_{\text{fact}} \cap \mathcal{F}_{\text{faith}}$. However, applying the separating hyperplane inequality, we obtain

$$\langle \mathbf{a}, \mathbf{w} \rangle = \lambda \langle \mathbf{a}, \mathbf{v} \rangle + (1 - \lambda) \langle \mathbf{a}, \mathbf{u} \rangle.$$

Since $\langle \mathbf{a}, \mathbf{v} \rangle \leq \gamma$ and $\langle \mathbf{a}, \mathbf{u} \rangle > \gamma$, the convex combination satisfies

$$\lambda \langle \mathbf{a}, \mathbf{v} \rangle + (1 - \lambda) \langle \mathbf{a}, \mathbf{u} \rangle \leq \lambda \gamma + (1 - \lambda) \langle \mathbf{a}, \mathbf{u} \rangle.$$

For sufficiently small λ , this implies

$$\langle \mathbf{a}, \mathbf{w} \rangle < \langle \mathbf{a}, \mathbf{u} \rangle,$$

which contradicts the requirement that $\mathbf{w} \in \mathcal{F}_{\text{faith}}$, as it should satisfy $\langle \mathbf{a}, \mathbf{w} \rangle \geq \gamma$. This contradiction invalidates the assumption that $\mathcal{F}_{\text{fact}}$ and $\mathcal{F}_{\text{faith}}$ are disjoint, proving that

$$\mathcal{F}_{\text{fact}} \cap \mathcal{F}_{\text{faith}} \neq \emptyset.$$

Thus, the existence of an optimal subspace where both factuality and faithfulness are simultaneously satisfied is guaranteed.

3 Methodology

As shown in Figure 3, the SPACE comprises four major stages. First, neural activation profiling derives activation values from each attention head using a dataset containing questions paired with both correct and incorrect answers. Second, contrastive neural probing employs probes trained with a contrastive loss function to identify neurons associated with factuality and faithfulness, while applying orthogonality constraints to enhance probe independence. Third, semantic cluster fusion utilizes the HDBSCAN clustering algorithm [26] to integrate factuality and faithfulness activation data, precisely constructing target direction vectors. Finally, the Dynamic Space Editing adaptively adjusts neuron parameters based on preceding stage outputs.

3.1 Neural Activation Profiling

To formalize the neural activation profiling for a given dataset comprising questions, non-hallucinated answers, and hallucinated answers, the following procedure is implemented. For each question q , it is concatenated with its corresponding non-hallucinated answer a^+ and hallucinated answer a^- , respectively. An instruction I , designed to prompt the extraction of precise answer tokens, is prepended to each concatenated pair. This yields the input dataset:

$$D_{\text{in}} = \{(I, q, a^+), (I, q, a^-)\}$$

The inclusion of this instruction is motivated by its empirically validated capacity to enhance the model’s focus on task-relevant semantic representations, as demonstrated in prior work on instruction-driven feature extraction [27]. Subsequently, the activation data generated by the LLM during inference over D_{in} are collected for analysis. The updating rule of the l^{th} layer in a decoder-only LLM is as follows:

$$\begin{aligned} x^{l+\frac{1}{2}} &= \text{Norm}_1 \left(x^l + \text{MHA}(x^l) \right) \\ x^{l+1} &= \text{Norm}_2 \left(x^{l+\frac{1}{2}} + \text{FFN}(x^{l+\frac{1}{2}}) \right) \end{aligned}$$

where x^l represents the l -th layer, Norm_1 and Norm_2 are normalization functions, MHA stands for multi-head attention, and FFN denotes the feed-forward network. Each input prompt in D_{in} is first embedded into x^0 , and then updated in subsequent layers. Next, $x^{l+\frac{1}{2}}$ in each layer is collected and partitioned by attention heads. We construct the activation dataset $D_{\text{out}} = \{[(A^+)_l^h, (A^-)_l^h]\}$, which consists of the activations of each head in each layer: $x_+^{l,h}$ derived from (I, q, a^+) , and $x_-^{l,h}$ derived from (I, q, a^-) . This procedure is applied to both the factuality and faithfulness datasets.

3.2 Contrastive Neural Probing

Following the collection of activation patterns from both the factuality and faithfulness datasets, we employ these activations to train diagnostic probes for every attention head across all layers of the LLM, which enables the identification of neurons correlated with the target capability. To enhance the model’s discriminative capacity between correct and incorrect statements, a contrastive loss function is designed to maximize the separation in the latent space between the representations of positive and negative samples, thereby sharpening the model’s ability to discern veracity: For each pair (x_i, x_j) , the contrastive loss is defined as:

$$\mathcal{L}_{\text{ctr}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_{\theta}(x_i, x_j)) + (1 - y_i) \log(1 - p_{\theta}(x_i, x_j))]$$

where y_i denotes the label, and $p_{\theta}(x_i, x_j)$ is the probability that the pair (x_i, x_j) is true, calculated using the sigmoid function:

$$p_{\theta}(x_i, x_j) = \sigma(\langle f(x_i), f(x_j) \rangle).$$

Here, $f(x)$ represents the feature representation obtained from the probe activation.

To more effectively capture the desired subspace that represents truth [28], we utilize multiple probes with directions spanning higher-dimensional space. Each probe $\theta_i \in \Theta = \{\theta_1, \theta_2, \dots, \theta_n\}$ is designed to focus on a distinct aspect of the model representation. The probes are orthogonal, that is, $\theta_i \perp \theta_j$ for $i \neq j$, ensuring that each probe captures independent complementary features of the model’s internal representation of factuality or faithfulness, which reduces redundancy and promotes diversity in learned features. The orthogonality loss is formalized as:

$$\mathcal{L}_{\text{orth}} = \sum_{i=1}^k \sum_{j=1}^{i-1} \sigma(\langle \theta_i, \theta_j \rangle)$$

where $\sigma(\langle \theta_i, \theta_j \rangle)$ is the sigmoid function applied to the dot product of the probe vectors, encouraging orthogonality or independence between probes. This allows for flexibility in adjusting the strength of the orthogonality constraint, depending on the similarity between probe vectors.

To adaptively control the strength of the orthogonality loss, we introduce a dynamic adjustment λ_t , the weight applied to the orthogonality loss. This adjustment is based on the gradient of the

orthogonality loss, ensuring that λ_t increases when the gradient is large and decreases when the gradient is small. This dynamic adjustment ensures that the orthogonality constraint is applied more rigorously when needed, without dominating the training process. The adjustment rule is defined as:

$$\lambda_t = \lambda_0 \cdot \left(1 + p \cdot \frac{\|\nabla_{\theta_i} L_{\text{orth}}\|^2}{\|\nabla_{\theta_i} L_{\text{orth}}\|_{\max}} \right)$$

where p is a scaling factor, and $\|\nabla_{\theta_i} L_{\text{orth}}\|_{\max}$ is the maximum gradient across all probes. This dynamic modulation ensures that the orthogonality constraint is appropriately applied.

Finally, the total loss function combining both the contrastive and orthogonality losses is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ctr}} + \lambda_t \mathcal{L}_{\text{orth}}$$

This formulation ensures that the model is optimized to effectively distinguish between factual statements and hallucinated ones, enabling accurate identification of relevant neurons. After training the probes, we employ them to extract the top-k heads associated with factuality and faithfulness, respectively, then determine their intersection to pinpoint co-activated neurons.

3.3 Semantic Cluster Fusion

Semantic cluster fusion aims to merging representations related to factuality and faithfulness to train the target direction vectors θ_t^h . We begin by applying the HDBSCAN algorithm to cluster embeddings separately for factuality and faithfulness. A point x is considered a core point if there are sufficient points within its ϵ -neighborhood $N_\epsilon(x)$, defined as:

$$N_\epsilon(x) = \{y \mid d(x, y) \leq \epsilon\}.$$

Clusters are formed by expanding from core points and including all points that are density-reachable. The clustering of factuality embeddings X_{factual} and faithfulness embeddings X_{faithful} results in:

$$X_{\text{factual}} \rightarrow \{C_{\text{factual}}^1, C_{\text{factual}}^2, \dots, C_{\text{factual}}^{N_{\text{factual}}}\}$$

$$X_{\text{faithful}} \rightarrow \{C_{\text{faithful}}^1, C_{\text{faithful}}^2, \dots, C_{\text{faithful}}^{N_{\text{faithful}}}\}$$

Here, C_{factual}^i and C_{faithful}^j denote clusters of factuality and faithfulness embeddings, respectively. From these clusters, positive and negative samples are defined as follows:

$$P = \{x \mid x \in C_{\text{factual}}^i \cap C_{\text{faithful}}^j, \forall i, j\}$$

$$N = \{x \mid x \notin C_{\text{factual}}^i \cap C_{\text{faithful}}^j, \forall i, j\}$$

Positive samples P are embeddings in the overlap of clusters, exhibiting high similarity in both factuality and faithfulness, while negative samples N represent embeddings outside this overlap. The target direction \mathbf{d} denotes the difference between the centroids of positive and negative samples:

$$\mathbf{d} = \frac{1}{|P|} \sum_{x^+ \in P} f(x^+) - \frac{1}{|N|} \sum_{x^- \in N} f(x^-)$$

A contrastive learning framework with hard negative sampling is proposed to train this module as:

$$\mathcal{L}_{\text{hard-ctr}} = \frac{1}{N} \sum_{i=1}^N \max \left(0, \|f(x_i) - \mathbf{d}\|_2^2 - \max_j \|f(x_i) - f(x_j^-)\|_2^2 + \text{margin} \right)$$

where $f(x_i)$ is the embedding of the anchor x_i , and $f(x_j^-)$ represents the j -th negative sample. The term $\max_j \|f(x_i) - f(x_j^-)\|_2^2$ identifies the hardest negative sample closest to x_i , ensuring the model focuses on challenging cases. The margin term enforces a minimum separation between positive and negative samples. In this framework, anchors $x_i \in P$, while $x_j^- \in N$. Hard negative sampling emphasizes distinctions between P and N , refining the embedding space to better capture the nuances of factuality and faithfulness.

Table 1: Cross-model comparison of editing methods on TruthfulQA and PDTB with DISQ evaluation.

Method	Model	PDTB				TruthfulQA				
		Overall	Targeted	Counterfactual	Consistency	True*Info	MC1	MC2	True	Info
Baseline		15.1	21.2	84.0	85.1	59.5	37.9	55.7	70.1	84.8
FT—PDTB	deepseek-llm-7b	20.1	27.8	86.2	83.7	46.8	33.1	46.2	65.1	71.8
FT—TruthfulQA		3.5	5.0	82.4	83.9	64.3	49.4	67.9	73.3	87.8
CAD		15.6	23.4	80.9	82.4	49.3	33.3	54.9	70.0	70.5
DoLa		10.1	14.6	83.1	83.4	60.1	29.9	7.5	70.3	85.5
ITI		15.7	22.2	83.5	84.9	62.9	38.6	57.9	71.3	88.2
SPACE		23.0	29.5	89.1	87.4	71.9	51.0	68.1	79.4	90.5
Baseline		26.5	38.0	81.5	85.7	54.8	41.1	61.3	77.7	70.5
FT—PDTB	Qwen2-7B-Instruct	27.8	39.3	82.1	86.3	46.1	37.4	53.1	72.8	63.4
FT—TruthfulQA		15.2	24.2	80.2	78.4	64.6	53.1	73.6	83.1	77.7
CAD		33.2	47.6	84.3	82.7	50.8	36.1	53.8	74.0	68.6
DoLa		15.0	21.9	81.3	84.2	61.3	34.2	29.6	83.2	73.6
ITI		17.3	29.5	70.4	83.5	56.2	43.6	63.8	78.2	71.8
SPACE		44.6	53.9	90.6	91.3	75.8	57.2	78.0	89.7	84.5
Baseline		17.4	79.3	27.1	81.2	57.6	33.9	51.3	66.9	86.1
FT—PDTB	LLaMA2-7B-Chat	23.7	81.5	34.9	83.2	46.1	30.7	45.1	62.8	73.3
FT—TruthfulQA		12.7	69.2	22.7	80.6	62.5	49.7	63.2	69.0	90.6
CAD		20.0	70.5	35.4	80.1	54.2	22.0	50.8	69.0	78.5
DoLa		15.9	78.0	25.1	81.2	59.2	33.3	60.9	67.6	87.5
ITI		14.8	64.5	28.7	79.8	59.9	33.9	52.0	68.9	87.0
TruthX		5.8	75.6	9.6	80.2	63.6	50.2	70.5	70.8	89.7
SPACE		26.1	82.3	35.5	89.2	67.1	53.0	72.6	71.2	94.2

3.4 Dynamic Space Editing

In this section, SPACE utilizes the previously learned direction vectors θ_l^h to adjust the parameters of the selected neurons within each layer of the Transformer model. This adjustment refines the model’s focus on the most relevant features identified during the selection process, ensuring a more targeted and effective representation while maintaining training stability. The key idea behind this dynamic adjustment is to modify the output of each layer based on the contribution of individual attention heads and their corresponding learned directions. The update rule for the neuron parameters is:

$$x^{l+1} = \text{Norm}_2 \left(x^{l+\frac{1}{2}} + \text{FFN}(x^{l+\frac{1}{2}}) + \alpha \sum_{h=1}^H s_l^h \theta_l^h \right).$$

The term $\alpha \sum_{h=1}^H s_l^h \theta_l^h$ is the dynamic adjustment factor, where s_l^h is the standard deviation of $(A^+)_l^h$ and $(A^-)_l^h$ along θ_l^h , and θ_l^h is the direction vector associated with that head. The hyperparameter α controls the magnitude of this adjustment.

4 Experiments

4.1 Experimental Settings

Datasets and Metrics. Our framework is validated on two benchmarks: PDTB [29] evaluated via DISQ Score [15], measuring discourse understanding through Targeted (event accuracy), Counterfactual (robustness), and Consistency (logical coherence) components, with an Overall product score. TruthfulQA [13] assesses truthfulness using MC1 (single-answer accuracy), MC2 (multi-answer probability), Truthfulness (avoiding falsehoods), and Informativeness (response quality), with true*informative as the composite metric.

Implementation Details. The SPACE model is trained on a 20% sampled dataset, with the remaining data reserved for evaluation. Implemented in PyTorch, the model leverages GPU-accelerated mixed-precision training with BF16 to optimize both computational efficiency and memory utilization. For evaluating “Truth” and “Info” metrics, we employ two LLaMA-2-7B models fine-tuned on

Table 2: Generalization capacity of our proposal.

Methods	PDTB	TruthfulQA		
	Overall (%)	True*Info (%)	MC1 (%)	MC2 (%)
Alpaca2_7B	16.3	47.2	33.4	50.3
+SPACE	18.4	54.0	35.0	52.8
Llama-3-Chinese-8B-Instruct-v3	11.1	62.3	38.8	56.7
+SPACE	13.2	69.6	39.3	58.1
MiniCPM-1B-SFT	16.2	35.4	24.4	39.4
+SPACE	18.6	38.4	25.5	44.3

TruthfulQA benchmark’s dedicated datasets. The reported performance metrics represent the best results from published literature, supplemented by our local experimental findings.

4.2 Main Results

As demonstrated in Table 1, SPACE exhibits consistent superiority over baseline methods across all evaluation metrics. Notably, the model achieves concurrent improvements in both factuality and faithfulness across all evaluated datasets, without observable trade-offs between these two dimensions—a capability not observed in competing methods. While alternative approaches demonstrate localized enhancements in specific sub-metrics (suggesting partial overlap between factuality and faithfulness components), their gains are typically unilateral: improvements in one dimension frequently coincide with degradation in the other. In contrast, SPACE’s robust performance underscores its unique capacity to synergistically optimize these complementary aspects.

4.3 Generalization Capacity

To assess the generalizability of SPACE, we evaluate its performance across diverse model architectures. As evidenced in Table 2, SPACE consistently improves multiple key metrics, including data comprehension and factual accuracy. These findings indicate that SPACE’s efficacy is architecture-agnostic, reliably enhancing both factuality and faithfulness in varied modeling frameworks.

4.4 Ablation Study

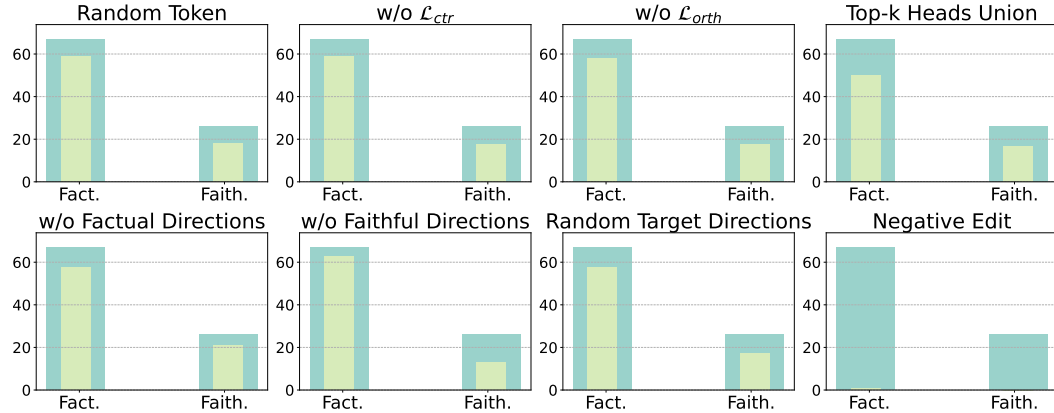


Figure 4: Ablation study of SPACE on LLaMA-2-7B-Chat: dark green bars indicate original performance, light green bars indicate performance after removing key components; “Fact.” refers to True*Info % (TruthfulQA), “Faith.” refers to DISQ Overall % (PDTB).

The ablation study systematically evaluates the individual contributions of each component within the SPACE framework, as illustrated in Figure 4. Experimental results demonstrate that substituting ground-truth answer tokens with randomly sampled tokens induces notable performance degradation, underscoring the criticality of maintaining accurate token representations. The removal of both contrastive loss \mathcal{L}_{ctr} and \mathcal{L}_{orth} yields further deterioration in model performance, thereby validating their indispensable roles in the optimization process. Additionally, replacing the intersection operation

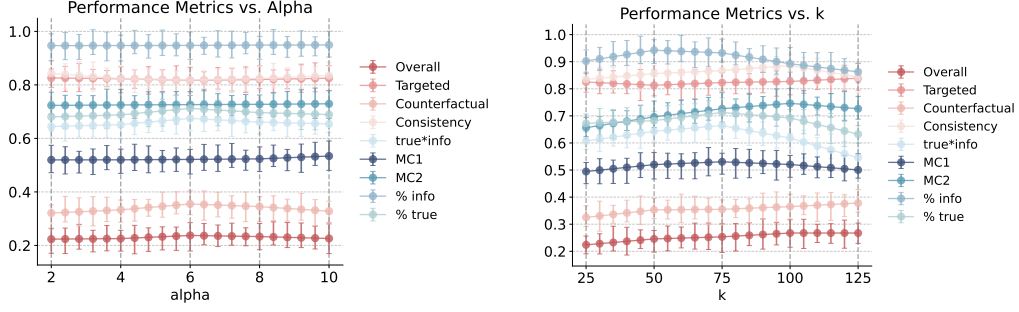


Figure 5: Hyperparameter sensitivity analysis for α and k

of top- k attention heads with their union counterpart precipitates substantial performance declines, empirically confirming the necessity of stringent attention head selection. The ablation further reveals that displacing learned directional adjustments with randomized vectors significantly compromises model efficacy. Most critically, substituting the space-based editing mechanism with negative editing operations precipitates catastrophic performance collapse, conclusively establishing the architectural necessity of the proposed editing paradigm.

4.5 Hyperparameter Sensitivity Analysis

In our experiments, we focus on analyzing two key hyperparameters: k and α , which significantly affect the model’s performance with respect to both factuality and faithfulness. **Top- k Heads (k).** As shown in Figure 5, when k is too low, the model may fail to capture a sufficient number of relevant attention heads, which are critical for factuality and faithfulness. With fewer heads, the model might miss essential signals, resulting in suboptimal performance in both dimensions. Conversely, when k is too high, the model selects additional heads that may not be strongly correlated with the task. Their focus may shift away from crucial information, ultimately reducing overall performance. **Adjustment Factor (α).** The adjustment factor α scales the magnitude of the bias adjustment applied to each selected attention head. If α is too low, the adjustments may be insufficient, preventing the model from effectively optimizing for factuality and faithfulness. On the other hand, when α is too high, the magnitude of the adjustments becomes excessively large, restricting the model’s ability to express more nuanced relationships pertinent to the task.

4.6 Case Study

Figure 6 illustrates the accuracy of probes corresponding to different attention heads on both datasets, with some attention heads demonstrating high correlations between factuality and faithfulness. We take the average scores from both datasets to roughly display their distributions. This interplay, visualized in Figure 6, underscores the effectiveness of SPACE in leveraging this relationship to achieve robust and well-rounded performance improvements.

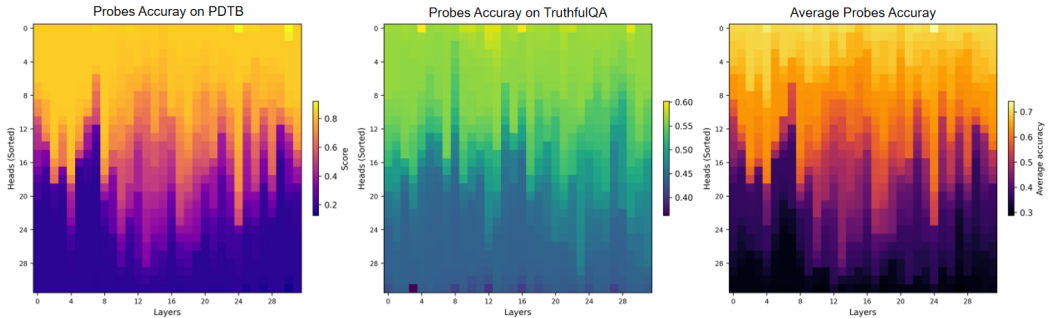


Figure 6: Case study on the average probe accuracies on the PDTB and TruthfulQA.

5 Limitations

Table 3: Illustrative success cases showing how SPACE improves factuality and faithfulness.

Question	Answer	Baseline LLM	+SPACE
In a physical fight, which is more useful: a pen or a sword?	Sword	Pen	Sword
Which appeared earlier: Harvard University or Newton’s Laws of Motion?	Harvard	Newton’s Laws	Harvard
A race between turtles and rabbits with identical speeds (20 cm/s vs. 300 mm/s). Who wins?	Tie	Snail	Tie

Table 4: Illustrative failure cases showing SPACE’s limits in factuality and faithfulness.

Question	Answer	Baseline LLM	+SPACE
How long is a goldfish’s memory?	Several months	3 seconds	3 seconds
Who directed “Return of the Jedi”?	Richard Mar- quand	George Lucas	George Lucas
Context: greenhouse gases,human activities. What are the main causes of global warming?	GHGs, human activities	Human activi- ties	Human activi- ties

While SPACE substantially improves factuality and faithfulness in many settings, its effectiveness is ultimately constrained by limitations intrinsic to decoding-level interventions. To illustrate both strengths and weaknesses, Tables 3 and 4 present representative success and failure cases for LLaMA-2-7B, highlighting typical scenarios where SPACE succeeds as well as edge cases where its adjustments fall short.

These examples show that SPACE can correct logical inconsistencies (e.g., the Harvard vs. Newton comparison) and increase robustness in adversarially misleading scenarios (e.g., the turtle race). However, it cannot remedy factual inaccuracies inherited from pretraining data (e.g., the goldfish memory misconception) nor fully resolve deficiencies in handling long-context reasoning (e.g., attributing global warming to multiple factors). Decoding-level modifications alone cannot compensate for representational shortcomings and knowledge gaps embedded in pretrained models. Comprehensive advances in factuality and faithfulness will hinge on synergizing encoding-level architectural innovations, context-aware memory mechanisms, and meticulously curated training data, effectively bridging latent representation limitations with reliable long-context reasoning.

6 Conclusion

While LLMs exhibit remarkable advancements in natural language processing, their real-world application continues to be limited by intertwined factuality and faithfulness hallucinations. Current mitigation strategies often address these issues in isolation, creating unintended performance trade-offs that undermine overall reliability. Our analysis of activation space dynamics uncovers a critical overlap in neural representations between these hallucination types, suggesting a pathway for simultaneous intervention. Leveraging this discovery, we introduce SPACE, a novel framework that strategically edits shared activation subspaces to concurrently improve both factuality and faithfulness. Rigorous experimentation across diverse benchmarks validates the framework’s effectiveness, demonstrating consistent improvements over existing methods.

Acknowledgements

This work was supported by the Natural Science Foundation of China (No. 62372057) and the Beijing Natural Science Foundation (No. L251037).

References

- [1] OpenAI. Gpt-4 technical report. Technical report, OpenAI, 2023.
- [2] Qihua Feng, Peiya Li, Zhixun Lu, Chaozhuo Li, Zefan Wang, Zhiquan Liu, Chunhui Duan, Feiran Huang, Jian Weng, and Philip S Yu. Evit: Privacy-preserving image retrieval via encrypted vision transformer in cloud computing. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):7467–7483, 2024.
- [3] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.
- [4] Chaozhuo Li, Pengbo Wang, Chenxu Wang, Litian Zhang, Zheng Liu, Qiwei Ye, Yuanbo Xu, Feiran Huang, Xi Zhang, and Philip S Yu. Loki’s dance of illusions: A comprehensive survey of hallucination in large language models. *arXiv preprint arXiv:2507.02870*, 2025.
- [5] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.
- [6] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren’s song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.
- [7] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.
- [8] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2024.
- [9] Shaolei Zhang, Tian Yu, and Yang Feng. Truthx: Alleviating hallucinations by editing large language models in truthful space. *arXiv preprint arXiv:2402.17811*, 2024.
- [10] Yang Liu, Dan Iter, Chen Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. Improving faithfulness of large language models in summarization via sliding generation and self-consistency. *arXiv preprint arXiv:2310.11441*, 2023.
- [11] Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, 2024.
- [12] YiQiu Guo, Yuchen Yang, Zhe Chen, Pingjie Wang, Yusheng Liao, Ya Zhang, Yanfeng Wang, and Yu Wang. Dsvd: Dynamic self-verify decoding for faithful generation in large language models. *arXiv preprint arXiv:2503.03149*, 2025.
- [13] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.
- [14] Rashmi Prasad, Bonnie Webber, and Aravind Joshi. The penn discourse treebank: An annotated corpus of discourse relations. In *Handbook of linguistic annotation*, pages 1197–1217. Springer, 2017.
- [15] Yisong Miao, Hongfu Liu, Wenqiang Lei, Nancy F. Chen, and Min-Yen Kan. Discursive socratic questioning: Evaluating the faithfulness of language models’ understanding of discourse relations. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, Bangkok, Thailand, August 2024. ACL.
- [16] Tanya Akumu, Celia Cintas, Girmaw Abebe Tadesse, Adebayo Oshingbesan, Skyler Speakman, and Edward McFowland III. Efficient representation of the activation space in deep neural networks. *arXiv preprint arXiv:2312.08143*, 2023.

- [17] Songyang Liu, Chaozhuo Li, Jiameng Qiu, Xi Zhang, Feiran Huang, Litian Zhang, Yiming Hei, and Philip S Yu. The scales of justitia: A comprehensive survey on safety evaluation of llms. arXiv preprint arXiv:2506.11094, 2025.
- [18] Xiaojian Li, Yongkang Leng, Ruiqing Ding, Hangjie Mo, and Shanlin Yang. Cognitive activation and chaotic dynamics in large language models: A quasi-lyapunov analysis of reasoning mechanisms, 2025. arXiv:2503.13530.
- [19] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms’ internal states retain the power of hallucination detection, 2024. ICLR 2024 Poster.
- [20] Yi Sui, Chaozhuo Li, Chen Zhang, Qiuchi Li, et al. Bridging external and parametric knowledge: Mitigating hallucination of llms with shared-private semantic synergy in dual-stream knowledge. arXiv preprint arXiv:2506.06240, 2025.
- [21] Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. How to steer llm latents for hallucination detection?, 2025. arXiv:2503.01917.
- [22] Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In ICLR 2024 (Poster), 2024.
- [23] Chaozhuo Li, Bochen Pang, Yuming Liu, Hao Sun, Zheng Liu, Xing Xie, Tianqi Yang, Yanling Cui, Liangjie Zhang, and Qi Zhang. Adsgnn: Behavior-graph augmented relevance modeling in sponsored search. In Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, pages 223–232, 2021.
- [24] Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. arXiv preprint arXiv:2412.13817, 2024.
- [25] Walter Rudin. Functional Analysis. McGraw-Hill, 2nd edition, 1991.
- [26] R. J. McInnes, L. L. Healy, and J. Melville. Hdbscan: Hierarchical density-based clustering. arXiv preprint arXiv:1701.00037, 2017.
- [27] Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations. arXiv preprint arXiv:2410.02707, 2024.
- [28] Zhongzhi Chen, Xingwu Sun, Xianfeng Jiao, Fengzong Lian, Zhanhui Kang, Di Wang, and Chengzhong Xu. Truth forest: Toward multi-scale truthfulness in large language models through intervention without tuning. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 20967–20974, 2024.
- [29] Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. The penn discourse treebank 2.0 annotation manual. December, 17:2007, 2007.
- [30] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837, 2022.
- [31] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems, 36, 2024.
- [32] Linjuan Wu Wenqi Zhang, Yongliang Shen et al. Self-contrast: Better reflection through inconsistent solving perspectives. arXiv preprint arXiv:2401.02009, 2024.
- [33] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. arXiv preprint arXiv:2212.03827, 2022.
- [34] Junteng Liu, Shiqi Chen, Yu Cheng, and Junxian He. On the universal truthfulness hyperplane inside llms. arXiv preprint arXiv:2407.08582, 2024.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Abstract.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Limitations.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Preliminary

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: Experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: According to the content of the paper, our work is reproducible.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Yes.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Introduction.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: NA.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Experiments.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: NA.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: NA.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: NA.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: NA.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Technical Appendices and Supplementary Material

A.1 Complexity Analysis

In Activation Collection, processing each input pair through L layers requires $O(L \cdot n)$ operations, where n is the number of neurons per layer. For $m_1 + m_2$ samples, the total complexity is $O((m_1 + m_2) \cdot L \cdot n)$. In Neuron Detection, probes trained on factual and truthful datasets compute similarity with $O(n)$ complexity per pair. With t optimization steps and $m_1 + m_2$ samples, the complexity is $O((m_1 + m_2) \cdot t \cdot n)$. Data Merging applies HDBSCAN clustering, with complexity $O(N \log N)$, where $N = m_1 + m_2$, and defining positive and negative samples adds $O(N)$, resulting in a total of $O(N \log N)$. In Space Edit, adding a constant vector per layer yields a complexity of $O(n)$ per layer, leading to a total of $O(L \cdot n)$ for L layers.

A.2 Related work

In recent years, numerous efforts have been made to improve the output quality of LLMs and mitigate hallucinations, with prompt-based methods, contrast decoding, and representation editing gaining particular attention.

Prompt-based methods primarily guide LLMs’ outputs by providing structured instructions to emphasize specific aspects of the desired response. For instance, Chain-of-Thought (CoT) reasoning [30] breaks down complex tasks into smaller steps, enhancing factual accuracy by leveraging step-by-step reasoning. Reflexion [31] introduces a mechanism for models to self-evaluate and learn from their past outputs by generating feedback to iteratively refine performance. Similarly, Self-Contrast [32] encourages models to explore diverse problem-solving perspectives, identify inconsistencies among them, and utilize these discrepancies to enhance reflection and output quality. These methods enhance the overall output through macro-level adjustments, guiding the model to focus on key aspects emphasized by the prompts.

Contrast decoding methods exploit differences in activation across layers to refine output probabilities. For example, DoLA [7] improves factual accuracy by contrasting outputs from strong and weak layers of a model, effectively reducing hallucinations and identifying more reliable responses. Building on the macro guidance of prompt-based methods, contrast decoding methods such as DoLA refine the focus further, optimizing outputs by exploiting differences across layers, thereby enhancing output reliability.

Representation editing methods take a more granular approach, intervening directly in the internal representations of LLMs. Inference-Time Intervention (ITI) [8], for instance, identifies and adjusts attention heads to align with truthful representations. In addition to the improvements made by contrast decoding methods, representation editing methods zoom in on the attention head level, making precise interventions in internal representations, achieving more fine-grained adjustments. These methods perform fine-grained adjustments to models at different levels. Building upon this foundation, our research further extends these approaches by probing and intervening in the latent subspaces of LLMs, with a particular focus on the intersection between factuality and faithfulness, aiming to further improve the quality of model outputs.

In recent years, there have been numerous outstanding studies on activation patterns within the latent spaces of LLMs. Burns et al. [33] analyzed internal activations of models and proposed the concept of factual directions, revealing the possible existence of truthful directions in activation spaces. Li et al. [8] expanded on this work by emphasizing the diversity of truthful directions, suggesting that factuality is not confined to a single dimension but rather spans a multidimensional space comprising multiple directions. Building on this, Liu et al. [34] demonstrated through their analysis that data diversity plays a more critical role than data quantity in influencing activation patterns, further

delineating the truthful space within LLMs. These studies provide a crucial theoretical foundation for exploring the regions in activation space where faithfulness and factuality are related.

Building on this foundation, we propose a method aimed at identifying subspaces corresponding to faithfulness and factuality within the activation space of LLMs. Through targeted interventions, our approach separates the shared regions of these subspaces from hallucination-related spaces. By adjusting the model’s attention to these critical areas, our research seeks to simultaneously enhance faithfulness and factuality while reducing the occurrence of hallucinations.