A DUAL SCALAR FIELD ON MANIFOLD FOR RECURSIVE H-NEURONS DETECTION AND MITIGATION IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucination-Associated Neurons (H-Neurons) refer to a pro-hallucinatory subset of separable neurons. Current methods identify this subset within the activation space by deploying linear probes to impose external statistical boundaries. Such an external empirical estimator lacks geometric interpretability, fundamentally struggling to yield a unified hallucination mitigation framework, while the exhaustive fitting of probes across successive layers incurs prohibitive cross-layer computational costs. Accordingly, we dually derive a scalar field on the manifold and recursively propagate it through the network layers, with its geometric properties intrinsically dictating H-Neurons detection and the mitigation direction. Specifically, we theoretically prove the equivalence between the tangent vector of the scalar field and that of the geodesic on the support connecting hallucinated states to truthful states. This fundamental equivalence enables both the identification of H-Neurons and the comparison of hallucination degrees: (1) decomposing the tangent vector onto activation axes allows component polarities to identify H-Neurons by revealing their pro-hallucinatory dynamical roles and (2) differentiating along the geodesic yields a non-negative derivative, establishing that the scalar field is monotonically non-decreasing along the geodesic trajectory toward truthful states. Subsequently, guided by tangential equivalence and monotonicity of the scalar field, we formalize hallucination mitigation as a path integral process governed by a first-order ordinary differential equation. This motivates a local trajectory along which hallucinated activations follow the ascent direction of the scalar field, thereby steering them toward regions associated with truthfulness. Finally, to alleviate cross-layer computational overhead, we introduce a manifold pullback mechanism to recursively derive high-quality scalar field candidates with a bounded approximation error, thereby reducing the need for exhaustive layer-wise fitting. Experiments demonstrate that our framework can accurately detect H-Neurons and significantly mitigate hallucinations in large language models.

1 INTRODUCTION

The exceptional utility of large language models (LLMs) in downstream tasks Chowdhery et al. (2023); Hoffmann et al. (2022) is continually undermined by the persistent challenge of hallucinations Deng et al. (2026); Ji et al. (2023); Lin et al. (2022); Manakul et al. (2023); Kuhn et al. (2023). To uncover the computational mechanisms underlying hallucinations, several studies shift away from viewing LLMs as macroscopic "black boxes," instead treating them as complex networks composed of microscopic computational units Dai et al. (2022); Geva et al. (2021); Meng et al. (2022). At this microscopic level, complex generative behaviors are fundamentally encoded within the representations of these specific neurons Meng et al. (2022). By examining the linear separability of hidden states Azaria & Mitchell (2023); Chen et al. (2024a) and neuron activation patterns Gao et al. (2025), researchers are striving to establish logical connections between macroscopic behavioral errors and specific neural circuits.

Pioneering this direction, recent studies Gao et al. (2025); Li et al. (2023) discovered Hallucination-Associated Neurons (H-Neurons) Dai et al. (2022): an extremely sparse subset within feed-forward networks Geva et al. (2021) that exacerbate the model's propensity for hallucinations Azaria &

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Mitchell (2023); Chen et al. (2024a). By forcibly fitting sparse linear probes Alain & Bengio (2018) over the activation space to derive a crude linear approximation of the dichotomy between truthful and hallucinated states, researchers successfully pinpointed these H-Neurons and established a causal link to model behavior Meng et al. (2022; 2023); Vig et al. (2020). Empirically, amplifying H-Neuron activations induces hallucinations, whereas suppressing them effectively mitigates such risks Li et al. (2023); Zou et al. (2024; 2023). Despite this empirical success, when used to identify H-Neurons, linear probing Belinkov (2022) primarily serves as an external empirical estimator that distinguishes hallucination activation through rigid statistical boundaries Alain & Bengio (2018); Li et al. (2023); Elazar et al. (2021), rather than characterizing the intrinsic dynamical roles of neurons in either promoting truthfulness or driving hallucinations, and therefore offers limited geometric interpretability Hewitt & Liang (2019). From a mitigation perspective, this linear dichotomy separates activation according to decision boundaries, but struggles to provide a mitigation trajectory. This gap fundamentally decouples hallucination detection from mitigation, forcing intervention methods to rely on heuristic Euclidean shifts with limited reliability Turner et al. (2024); Marks & Tegmark (2024); Canby et al. (2025). Furthermore, because activation distributions shift across network depths, probes must be exhaustively fitted to each successive layer, incurring prohibitive computational costs Orgad et al. (2025); Gao et al. (2025); Zou et al. (2023).

To mitigate the aforementioned challenges, we dually derive an intrinsic scalar field on the manifold and recursively propagate it across network layers, utilizing its geometric properties to unify H-Neuron detection and the mitigation direction. Specifically, building upon the Manifold Hypothesis of LLM representations Kari (2025); Tiblias et al. (2026), intermediate activation spaces are modeled as Riemannian manifolds with distinct truthful and hallucinated activation distributions. To quantify their geometric discrepancy, the Wasserstein-1 distance is formulated through Kantorovich-Rubinstein duality Villani (2008), which yields an optimal 1-Lipschitz scalar field. We mathematically prove that, on the support where the duality saturation condition holds, applying the inverse metric tensor to the covariant derivative of this scalar field yields a tangent vector that coincides with the normalized tangent vector of the geodesic connecting the hallucinated and truthful states. This fundamental equivalence leads to two key corollaries: (1) First, the scalar field natively facilitates structural neuron decoupling: when the tangent vector is decomposed onto specific neuron activation axes, its component values characterize the neuron's dynamical role along the geodesic, whereby positive components indicate a positive correlation with truthfulness, and negative components define the H-Neurons. (2) Second, the scalar field enables a precise comparison of hallucination levels, as differentiating the field along the geodesic path from hallucinated to truthful states yields a non-negative derivative, which tends to ensure that the scalar value is monotonically non-decreasing along the geodesic toward truthfulness. Subsequently, predicated on the tangential equivalence and field monotonicity, the computation of global geodesics is bypassed. Instead, hallucination mitigation is formalized as a path integral process governed by a local first-order ordinary differential equation (ODE). This establishes an ascending trajectory of scalar field that inherently amplifies truthfulness, allowing the tangent vector to pointwisely provide a principled local direction for steering hallucination activations toward truthfulness. Finally, to alleviate the computational burden across layers, we recursively derive approximate scalar field candidates across activation spaces using a regularization operator grounded in the manifold pullback mechanism, thereby circumventing the prohibitive requirement for exhaustive layer-wise fitting. We show that the recursive error bound is governed by the structural constants of the network, providing a controllable upper bound on the approximation error across layers. Extensive experiments validate our theoretical claims and derived corollaries, demonstrating that our approach enables effective H-Neuron identification and significantly mitigates hallucinations in LLMs. Our main contributions are four-fold:

- By leveraging duality, we derive an intrinsic scalar field and prove that the tangent vector of this scalar field is equivalent to the tangent vector of the geodesic. We further derive the polarity correspondence between H-Neurons and tangent vector components, while establishing the monotonicity of the scalar field with respect to truthfulness.
- Given the tangential equivalence and field monotonicity, hallucination mitigation can be formalized as a path integral governed by a first-order ordinary differential equation, which prescribes the ascending trajectory for reverting from a hallucinated state to a truthful state.
- To reduce the computational overhead of scalar field solving across layers, we leverage a normalized manifold pullback operator to recursively derive high-quality scalar field candidates, with error bounds controlled by network structural constants.

- Building upon these theoretical findings, we develop a principled framework driven by intrinsic scalar fields on activation manifolds. This framework provides a geometric perspective for characterizing H-Neurons and guiding hallucination mitigation. Extensive experiments further provide empirical support for our theoretical analysis and demonstrate the effectiveness of the framework in H-Neuron detection and hallucination mitigation.

2 PRELIMINARIES

Assumption 1 (**Local Manifold Assumption for Activation Distribution**). *Grounded in the Manifold Hypothesis of LLMs Kari (2025); Tiblias et al. (2026), the intermediate activation states are locally concentrated on a smooth d-dimensional differentiable manifold $\mathcal{M}_L$.*

Definition 1 (**Riemannian Structure of the Activation Manifold**). *Under Assumption 1, we define the intermediate activation manifold as a Riemannian manifold $(\mathcal{M}_L, g_{bc})$.*

- *$\mathcal{M}_L$ is a d-dimensional differentiable manifold satisfying $\mathcal{M}_L = \bigcup_\alpha O_\alpha$, where each O_α is an open set of $\mathcal{M}_L$. For each O_α, there exists a local chart $\psi_\alpha : O_\alpha \to \psi_\alpha(O_\alpha) \subset \mathbb{R}^d$. For any activation state $p \in O_\alpha$, its local coordinate representation is $\psi_\alpha(p) = \{x^\mu\} = (x^1, \ldots, x^\mu, \ldots, x^d)$, where x^μ denotes the activation value of the μ-th neuron.*
- *g_{bc} is a symmetric positive-definite metric tensor field on $\mathcal{M}_L$, which defines the inner product and norm of tangent vectors. For any $u, v \in T_p\mathcal{M}_L$, the induced inner product and norm are $\langle u, v\rangle_g = g_{bc}u^b v^c$ and $\|v\|_g = \sqrt{g_{bc}v^b v^c}$, respectively. For any smooth curve $\gamma : [0, \ell] \to \mathcal{M}_L$, its length is defined as $L_g(\gamma) = \int_0^\ell \|\dot\gamma(t)\|_g dt$, and the geodesic distance between $p, q \in \mathcal{M}_L$ is given by $d_g(p, q) = \inf_{\gamma(0)=p,\gamma(\ell)=q} L_g(\gamma)$, where $\dot\gamma(t)$ denotes the tangent vector along the curve.*

We adopt a Fisher matrix for the metric tensor (Appendix B.1), with additional geometric definitions and operations provided in Appendices B and B.2, together with Lemmas 1, 2, and 3.

3 HALLUCINATION NEURONS DECOUPLING AND MITIGATION

3.1 DETECTION OF H-NEURONS BASED ON A KANTOROVICH-RUBINSTEIN SCALAR FIELD

Recent studies Gao et al. (2025); Li et al. (2023) have confirmed the existence of H-Neurons and further shown that amplifying their activations increases hallucination-related behaviors. However, existing identification methods based on linear probes, while empirically effective, provide only limited insight into the intrinsic geometric mechanism underlying hallucination. We therefore seek a more robust theoretical characterization. Specifically, we introduce an intrinsic scalar field on $(\mathcal{M}_L, g_{ab})$, through which hallucination directions can be identified directly from the internal geometry of the activation space, while mitigation strategies can also be derived accordingly.

Let μ_L^- and μ_L^+ represent the empirical distributions of hallucinated and truthful activation states, respectively. To quantify the mitigation effort, we employ the Wasserstein-1 distance. In its primal formulation, it seeks an optimal transport plan π^* minimizing the expected geodesic cost:

$$W_1(\mu_L^-, \mu_L^+) = \inf_{\pi \in \Pi(\mu_L^-, \mu_L^+)} \int_{\mathcal{M}_L \times \mathcal{M}_L} d_g(y, x)\, d\pi(y, x), \tag{1}$$

Here, $\Pi(\mu_L^-, \mu_L^+)$ denotes the set of joint distributions on $\mathcal{M}_L \times \mathcal{M}_L$ with marginals μ_L^- and μ_L^+, where $y \sim \mu_L^-$ and $x \sim \mu_L^+$ correspond to hallucinated and truthful states. By Kantorovich-Rubinstein duality, Eq. 1 equals the maximum expectation gap over 1-Lipschitz scalar fields:

$$W_1(\mu_L^+, \mu_L^-) = \sup_{f_L \in \mathrm{Lip}_1(\mathcal{M}_L)} \left[\int_{\mathcal{M}_L} f_L\, d\mu_L^+ - \int_{\mathcal{M}_L} f_L\, d\mu_L^- \right] \tag{2}$$

Here, $f_L \in \mathrm{Lip}_1(\mathcal{M}_L)$ denotes a scalar field satisfying the unit Lipschitz constraint. For differentiable f_L, this reduces to the gradient constraint $|\nabla f_L|_g \le 1$:

$$\|\nabla f_L\|_g = \sqrt{g^{ab}(\nabla_a f_L)(\nabla_b f_L)} \le 1 \quad \text{a.e.} \tag{3}$$

The supremum in the duality is attained by the optimal scalar field, denoted as f_L^*. This scalar field is mathematically formulated as the argmax over the Lipschitz space:

$$f_L^* = \underset{f \in \mathrm{Lip}_1(\mathcal{M}_L)}{\arg\max} \left[\int_{\mathcal{M}_L} f(x) d\mu_L^+(x) - \int_{\mathcal{M}_L} f(x) d\mu_L^-(x) \right]. \tag{4}$$

Along optimal transport paths, the Lipschitz constraint is saturated, yielding for $(x, y) \in \mathrm{supp}(\pi^*)$

$$f_L^*(x) - f_L^*(y) = d_g(x, y). \tag{5}$$

A complete discussion of the Kantorovich–Rubinstein duality is provided in Appendix C.1. The practical training procedure of f_L^* is detailed in Appendix E.4. Latin indices on quantities such as $\nabla_a f$ and g^{ab} in Eq. 3 denote abstract indices, with calculation rules detailed in Appendix B.2.4.

Theorem 1 (**Equivalence Between Geodesics and Scalar Field Tangent Vectors**). *Let $\gamma : [0, \ell] \to \mathcal{M}_L$ be a geodesic connecting a hallucinated activation y and a truthful activation x, with $(y, x) \in \mathrm{supp}(\pi^*)$, where π^* denotes the optimal transport plan. According to the duality saturation in Eq. 5, together with the line integral theorem and geodesic length definition, we obtain:*

$$f_L^*(x) - f_L^*(y) = \int_0^\ell \left\langle g^{ab} \nabla_b f_L^*(\gamma(t)), \dot{\gamma}(t) \right\rangle_g dt = \int_0^\ell \|\dot{\gamma}(t)\|_g \, dt = d_g(x, y). \tag{6}$$

where $\dot{\gamma}(t)$ represents the tangent vector of the geodesic at the point $\gamma(t)$ with respect to the geodesic parameter t, i.e., $\dot{\gamma}(t) = \frac{d\gamma(t)}{dt}$. On the other hand, by the gradient-norm bound in Eq. 3 and the Cauchy–Schwarz inequality, the directional derivative along γ satisfies the pointwise upper bound

$$\left\langle g^{ab} \nabla_b f_L^*(\gamma(t)), \dot{\gamma}(t) \right\rangle_g \le \left\| g^{ab} \nabla_b f_L^*(\gamma(t)) \right\|_g \|\dot{\gamma}(t)\|_g \le \|\dot{\gamma}(t)\|_g. \tag{7}$$

Combining Eq. 6 and Eq. 7, the upper bound in Eq. 7 must be attained almost everywhere along γ; otherwise, the integral on the left-hand side of Eq. 6 would be strictly smaller than the total geodesic length. Therefore, the equality conditions of the Cauchy–Schwarz inequality and the gradient-norm bound hold almost everywhere, yielding

$$g^{ab} \nabla_b f_L^*(\gamma(t)) = \frac{\dot{\gamma}^a(t)}{\|\dot{\gamma}(t)\|_g} \qquad \text{for a.e. } t \in [0, \ell]. \tag{8}$$

According to Lemma 2, $g^{ab} \nabla_b f_L^*(\gamma(t))$ is the tangent vector v^a induced by f_L^*. The detailed proof and derivation are provided in Appendix C.2.

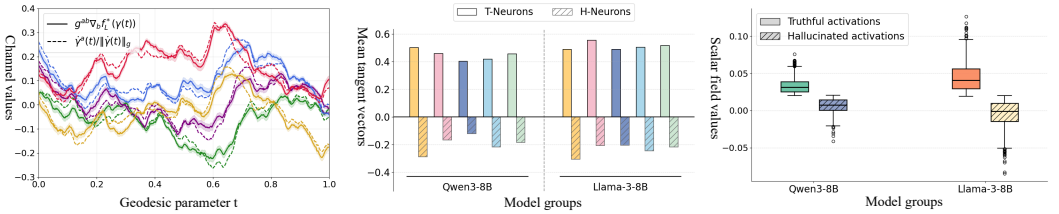
Figure 1(a) shows the alignment between $\dot{\gamma}^a(t)$ and $g^{ab} \nabla_b f_L^*(\gamma(t))$, while Appendix F.10 quantifies this alignment via cosine similarity, providing empirical support for Theorem 1.

Remark: According to the identity $g^{ab} \nabla_b f_L^*(\gamma(t)) = \frac{\dot{\gamma}^a(t)}{\|\dot{\gamma}(t)\|_g}$, the tangent vector of the scalar field exactly characterizes this transition direction along the geodesic. Consequently, once f_L^* is constructed, $g^{ab} \nabla_b f_L^*$ provides a computable intrinsic direction for characterizing the transition from hallucinated activation states toward the truthful region. At the neuron level, this direction further reveals the activation sensitivity of individual neurons during the transition from hallucination to truthfulness. In this sense, f_L^* provides both a geometric criterion for identifying H-Neurons and a principled direction for hallucination mitigation.

Corollary 1 (**Scalar Field Based Identification of H-Neurons**). *For a local activation state $\gamma(t)$ along the geodesic, since f_L^* is a smooth scalar field defined on the activation manifold, its covariant derivative can be expanded as $\nabla_a f_L^* = \sum_{\mu=1}^d (\nabla_\mu f_L^*)(dx^\mu)_a, \nabla_\mu f_L^* = \left(\frac{\partial}{\partial x^\mu}\right)^a \nabla_a f_L^* = \frac{\partial f_L^*}{\partial x^\mu}$. Applying the equality in Theorem 1 to the μ-th neuron activation, and combining it with the tangent vector computation in Lemma 2 and the preceding expansion, we obtain*

$$v^\mu(t) = g^{\mu\nu} \nabla_\nu f_L^*(\gamma(t)) = g^{\mu\nu} \frac{\partial f_L^*}{\partial x^\nu}(\gamma(t)) = \frac{1}{\|\dot{\gamma}(t)\|_g} \frac{d\gamma^\mu(t)}{dt}. \tag{9}$$

*Here, $v^\mu(t)$ denotes the μ-th component of the tangent direction induced by f_L^*. Since $\frac{d\gamma^\mu(t)}{dt} = \|\dot{\gamma}(t)\|_g v^\mu(t)$, and $\|\dot{\gamma}(t)\|_g > 0$ for a regular geodesic, **the sign of $v^\mu(t)$ directly determines whether the μ-th activation is locally enhanced or suppressed along the mitigation direction**. Detailed explanations and proofs can be found in the Appendix C.3.*



(a) The tangent vectors of f_L^* and $\gamma(t)$ (b) The tangent vectors across different neurons (c) Scalar field values of different distributions

Figure 1: Geometric analysis of the learned scalar field on TruthfulQA Lin et al. (2022). (a) Comparison of the normalized geodesic tangent vector $\dot{\gamma}^a(t)/\|\dot{\gamma}(t)\|_g$ and the scalar field tangent vector $g^{ab}\nabla_b f_L^*(\gamma(t))$ along the geodesic path, with five randomly selected components visualized. (b) Mean tangent vector components of H-Neurons and Truthfulness-Associated Neurons (T-Neurons). (c) Boxplot comparison of scalar field for samples from truthful and hallucinated distributions.

- ***Hallucination-Associated Neurons (H-Neurons)*** *($v^\mu = g^{\mu\nu}\nabla_\nu f_L^* = g^{\mu\nu}\frac{\partial f_L^*}{\partial x^\nu} < 0$): A negative local tangent component indicates that the μ-th activation is locally decreasing along the geodesic. This neuron corresponds to a hallucination-direction whose activation is locally suppressed during the transition toward the truthful state. A more comprehensive description of H-Neurons can be found in Definition 3.*

As shown in Figure 1(b), the mean tangent components over five samples across models are consistently negative for H-Neurons and positive for T-Neurons, further supporting the above analysis.

Corollary 2 (**Monotonicity of the Scalar Field along Geodesics**). *Let y denote a hallucinated state and x denote a truthful state. Consider the geodesic $\gamma : [0, \ell] \to \mathcal{M}_L$ satisfying $\gamma(0) = y$ and $\gamma(\ell) = x$. The monotonic increase can be further verified by differentiating f_L^* along the geodesic $\gamma(t)$. By the definition of the norm in Definition 1 and the equivalence in Theorem 1, we obtain*

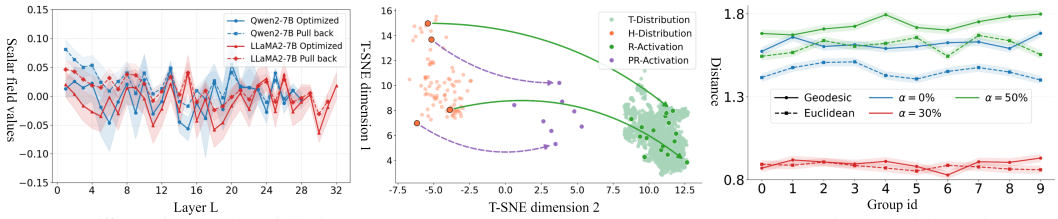
$$\frac{d}{dt} f_L^*(\gamma(t)) = \nabla_a f_L^* \frac{d\gamma^a}{dt} = \|\dot{\gamma}(t)\|_g \, \nabla_a f_L^* \left(g^{ab}\nabla_b f_L^*\right) = \|\dot{\gamma}(t)\|_g \, \|\nabla f_L^*\|_g^2 \geq 0. \tag{10}$$

The nonnegative derivative shows that f_L^* is non-decreasing along γ as the trajectory progresses from the hallucinated state y toward the truthful state x. ***Consequently, within this trajectory, a larger value of f_L^* indicates that the corresponding state is relatively more aligned with the truthful distribution.*** *Detailed explanations and proofs can be found in the Appendix C.4.*

Figure 1(c) shows that the truthful distribution consistently receives larger scalar field values than the hallucinated distribution, which further supports the monotonicity stated in Corollary 2.

Remark: Corollary 1 shows that the scalar field f_L^* identifies neuron-level activation trends through the tangent component $v^\mu = g^{\mu\nu}\nabla_\nu f_L^*$. Specifically, $v^\mu < 0$ means that the activation value of the μ-th neuron decreases along the transport trajectory from hallucination to truthfulness, indicating an H-Neuron whose activation is suppressed during the hallucination mitigation process. Corollary 2 further complements this local criterion with a scalar field ordering principle: differentiating f_L^* along the transport trajectory gives $\frac{d}{dt} f_L^*(\gamma(t)) \geq 0$, showing that the scalar field is non-decreasing from hallucinated states toward truthful states.

We emphasize that the tangent vector equivalence in Theorem 1 and the results in Corollaries 1 and 2 hold only along geodesics connecting optimal transport pairs in the support of π^* that satisfy the duality saturation condition. These conclusions are not guaranteed at arbitrary points of $\mathcal{M}_L$ or along arbitrary geodesics. In our empirical setting, we construct the scalar field from empirical measures of truthful and hallucinated activation samples. **As the sample size increases, these measures cover more relevant transport pairs and represent a larger portion of the support of π^*, expanding the region where duality saturation and the tangent vector equivalence can be empirically verified.** The saturation analysis in Appendix F.8 provides empirical evidence for this claim. We also assess robustness across sampling procedures in Proposition 3 and describe the training, detection, and remediation procedures in Algorithm 1.



(a) Difference between the optimized and the pullback scalar field.

(b) Repair status of hallucination related activations in Qwen2-7B.

(c) Distance Evolution under Different Intervention Ratios

Figure 2: (a) Difference between the optimized scalar field and the pullback scalar field across different layers in Qwen2-7B and Llama2-7B. (b) The T-Distribution and H-Distribution denote the truthful and hallucinated activation distributions. R-Activation indicates rectified activations that are successfully guided toward the truthful distribution, while PR-Activation denotes partially rectified activations that move away from the hallucinated region but remain insufficiently aligned with the truthful distribution. (c) Geodesic and Euclidean Distance Analysis under Different Intervention Ratios. Solid and dashed curves denote geodesic and Euclidean distances, respectively. Data used in these experiments are derived from TruthfulQA Lin et al. (2022).

3.2 Scalar field-guided path integral for hallucination mitigation

In practical hallucination mitigation, the truthful target corresponding to a hallucination-prone hidden activation is generally unknown. Therefore, instead of explicitly constructing a complete geodesic between a hallucinated state and a truthful state, we seek a local mitigation rule that only depends on the current activation state and the learned scalar field.

According to Theorem 1, we have $g^{ab}\nabla_b f_L^*(\gamma(t)) = \frac{\dot\gamma^a(t)}{\|\dot\gamma(t)\|_g}$. Here, $\dot\gamma^a(t)$ denotes the tangent vector pointing along the hallucination to truthfulness transport direction. Thus, the tangent vector $g^{ab}\nabla_b f_L^*$ provides a computable direction for rectifying hallucination-prone activations toward truthful regions. Consider an activation $H_L \in \mathcal{M}_L$ with local coordinates $(x^1, \dots, x^d)$, where $x^i = x^i(H_L)$. For the μ-th neuron, its activation coordinate evolves according to

$$\begin{cases} \dfrac{dx^\mu(t)}{dt} = g^{\mu\nu}(x(t))\dfrac{\partial f_L^*(x(t))}{\partial x^\nu}, & \mu = 1, \dots, d, \\ x^\mu(0) = x^\mu(H_L). \end{cases} \tag{11}$$

Under the existence–uniqueness conditions for first-order ODEs, the above-defined mitigation flow admits a unique local trajectory determined by the initial activation state H_L and the scalar field f_L^*.

This mitigation flow is consistent with the monotonicity established in Corollary 2. Specifically, Corollary 2 shows that the scalar field f_L^* is non-decreasing along the transport direction from hallucinated states toward truthful states. Since the mitigation flow uses the same tangent vector direction, it moves the current activation toward regions with larger scalar field value.

As shown in Figure 2(b), most hallucinated activations are shifted toward the truthful distribution, while partially rectified activations still follow the desired direction. Figure 2(c) further compares different intervention ratios across 10 sample groups, showing that a moderate intervention ratio (30%) achieves a notable reduction in distance to the truthful distribution compared with no intervention (0%), indicating the effectiveness of the proposed intervention mechanism. The ODE intervention procedure and optimization algorithm are detailed in Appendix E.5 and Algorithm 1.

Remark: The above construction shows that the final truthful distribution does not need to be explicitly specified. Given the initial activation and the equivalent condition in Theorem 1, the mitigation process can be formulated as a first-order ordinary differential system induced by the scalar field f_L^*. By the standard existence–uniqueness theorem, this system determines a unique local correction trajectory. This trajectory evolves along the direction induced by the geodesic. By Corollary 2, f_L^* increases monotonically along this direction, driving activations toward the truthful distribution and effectively transforming the computation of a complete geodesic into the pointwise determination of the f_L^* ascent direction.

4 PULLBACK OF SCALAR FIELDS ACROSS LAYERS

In the above section, we constructed a scalar field and theoretically showed that it serves as a unified geometric tool for hallucination detection and mitigation. Considering the large depth of LLMs, repeatedly optimizing the scalar field across different layers would introduce substantial computational overhead. To mitigate this issue, we further show that a high-quality scalar field at layer $L-1$ can be effectively induced from the scalar field at layer L through a pullback mechanism.

Assumption 2 (Local Lipschitz Continuity of Cross-Layer Mapping). *The cross-layer mapping $\phi_{L-1} : \mathcal{M}_{L-1} \to \mathcal{M}_L$, defined by the composition of residual updates and layer normalization in Definition 2, is assumed to be locally Lipschitz continuous with respect to the metrics $g^{(L-1)}$ and $g^{(L)}$. Specifically, for any compact subset $\mathcal{K} \subset \mathcal{M}_{L-1}$ (representing the bounded domain of valid hidden states), there exists a local constant $K_{L-1} > 0$ such that for any $H_x^{(L-1)}, H_y^{(L-1)} \in \mathcal{K}$,*

$$d_{g^{(L)}}\Big(\phi_{L-1}(H_x^{(L-1)}), \phi_{L-1}(H_y^{(L-1)})\Big) \le K_{L-1}\, d_{g^{(L-1)}}\Big(H_x^{(L-1)}, H_y^{(L-1)}\Big). \tag{12}$$

Proposition 1 (Recursive Construction via Normalized Pullback). *Consider the cross-layer mapping $\phi_{L-1} : \mathcal{M}_{L-1} \to \mathcal{M}_L$ as defined in Definition 2. Assume that ϕ_{L-1} satisfies the K_{L-1}-Lipschitz continuity in Assumption 2, and let f_L^* be the optimal 1-Lipschitz scalar field on $\mathcal{M}_L$. Then a candidate scalar field on $\mathcal{M}_{L-1}$ can be induced by the normalized pullback operator:*

$$f_{L-1}(H^{(L-1)}) = \frac{1}{K_{L-1}}(\phi_{L-1}^* f_L^*)(H^{(L-1)}) = \frac{1}{K_{L-1}}\, f_L^*\Big(\phi_{L-1}(H^{(L-1)})\Big), \tag{13}$$

where ϕ_{L-1}^* denotes the pullback operator that transports the truthfulness scalar field from layer L to layer $L-1$. In particular, the pullback satisfies $\phi_{L-1}^* f = f \circ \phi_{L-1}$.

To theoretically justify the effectiveness of this pullback mechanism, the following proposition establishes an explicit upper bound on the approximation error.

Proposition 2 (Cross-Layer Structural Suboptimality Bound). *Let $\phi_{L-1} : \mathcal{M}_{L-1} \to \mathcal{M}_L$, be the cross-layer mapping defined in Definition 2. The suboptimal error of the recursive candidate scalar field f_{L-1} is bounded by the layer-wise structural constants:*

$$\mathrm{Err}_{L-1} \le \Delta_{L-1}^* - \frac{\Delta_L^*}{\mathrm{Lip}(\sigma)\cdot s_1(W_{\text{in}}^{(L)})\cdot \mathrm{Lip}(\mathrm{LN})\cdot\big(1+\mathrm{Lip}(\mathrm{Attn}_L)\,\mathrm{Lip}(\mathrm{LN})\big)\cdot s_1(W_{\text{out}}^{(L-1)})}, \tag{14}$$

where $\Delta_{L-1}^* = W_1(\mu_{L-1}^+, \mu_{L-1}^-)$ denotes the Wasserstein-1 separation at layer $L-1$, and $s_1(\cdot)$ denotes the spectral norm of the corresponding weight matrices. Here, $W_{\text{out}}^{(L-1)}$ and $W_{\text{in}}^{(L)}$ correspond to the output transformation of layer $L-1$ and the input transformation of layer L, respectively. Notably, the entire denominator constitutes the analytical upper bound of the cross-layer local Lipschitz constant K_{L-1}. Detailed explanations and proofs can be found in Appendix F.11.

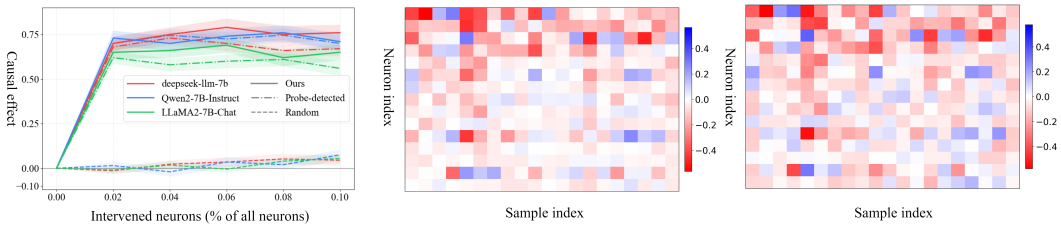
As shown in Figure 2(a), the optimized and pullback scalar fields exhibit consistent trends, providing empirical evidence for the controllable approximation error established in Proposition 2. Furthermore, detailed empirical statistics of the local Lipschitz constants are provided in Appendix F.11.

Remark: The above results show that a high-quality candidate scalar field at layer $L-1$ can be recursively induced from the scalar field at layer L, avoiding redundant optimization across layers. Together with the recursive metric propagation established in Lemma 3, the induced geometry provides the necessary operators for recursively applying the scalar field-guided hallucination detection and mitigation mechanism. The resulting structural suboptimality bound characterizes the gap between the induced and directly optimized scalar fields, governed by the structural constants of the cross-layer mapping.

5 EXPERIMENTAL VALIDATION

5.1 COMPARISON WITH BASELINES

As reported in Table 1, the proposed method achieves consistently superior performance over the baseline approaches across both PDTB Prasad et al. (2008) and TruthfulQA Lin et al. (2022) evalua-



(a) Causal effects of intervening hallucination neurons

(b) The 15 randomly selected neurons over the same batch of samples

(c) Probe-based scores at the same 15 neuron positions selected by our method

Figure 3: (a) Comparison of the causal effects of interventions on H-Neurons identified by our method, H-Neurons detected by linear probes, and randomly selected neurons across three LLMs. (b) Negative tangent vector components of 15 neurons randomly sampled from those identified by our method. (c) Negative probe scores of the same 15 neurons. Red denotes T-Neurons, while blue denotes H-Neurons. All experiments are conducted on TruthfulQA Lin et al. (2022).

tion metrics. This result provides empirical support for the scalar field-guided mitigation mechanism introduced in Section 3.2. Rather than relying on task-specific fine-tuning or explicitly specifying the final truthful target, our method uses the tangent vector of the learned scalar field as a local correction direction. Consequently, hallucination-prone activations are progressively transported toward regions associated with more truthful hidden states. A key observation is that our method improves factuality and faithfulness simultaneously across the evaluated datasets. In contrast, competing methods often show localized gains on individual sub-metrics, suggesting that they can partially improve either factual correctness or response faithfulness.

5.2 Causal Effects of Interventions

As shown in Figure 3, we analyze H-Neurons from both intervention effect and activation pattern perspectives. Figure 3(a) compares the neurons identified by our method with those selected by linear probes and random selection. The proposed method achieves the largest causal intervention effect, suggesting that the identified neurons are more closely associated with hallucination behaviors. Neurons selected by linear probes also exhibit effective intervention effects, but their performance is weaker than that of our method while remaining better than random selection.

5.3 Visualization of Hallucination Neurons

Figure 3(b) shows the selected neurons identified by our method, among which several neurons exhibit clear hallucination-associated activation patterns. Figure 3(c) presents the corresponding probe scores at the same neuron positions. The discrepancy between the two response patterns suggests that the scalar field-based criterion captures complementary hallucination-associated directions beyond conventional linear probe-based detection.

5.4 Bidirectional Patching for Causal Verification

Figure 4(a) reports the score changes under bidirectional activation patching. Patching hallucination-associated activations toward their truthful counterparts consistently improves performance across different models, whereas applying the reverse intervention leads to performance degradation. These results provide empirical evidence that the proposed scalar field-guided mechanism identifies directions associated with hallucination and can effectively steer them toward more truthful states.

5.5 Layer-wise Pullback Repair Analysis

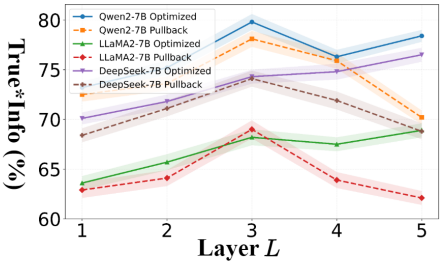
Figure 4(b) presents the layer-wise True*Info results of the optimized and pullback scalar fields on Qwen2-7B, LLaMA2-7B, and DeepSeek-7B. The pullback scalar field closely follows the optimized one in shallow layers, demonstrating the effectiveness of the recursive pullback mechanism. Although approximation errors accumulate with increasing depth, the pullback scalar field preserves similar improvement trends and achieves consistent truthfulness gains across models.

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
Ours		**26.4** (↑14.8%)	**31.3**(↑6.1%)	**90.7**(↑1.8%)	**88.9**(↑1.7%)	**74.1**(↑3.1%)	**56.2**(↑9.3%)	**70.0**(↑2.8%)	**82.5**(↑3.9%)	**91.3**(↑0.9%)
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
Ours		**46.2**(↑3.6%)	**55.7**(↑3.3%)	**90.8**(↑0.2%)	**93.1**(↑2.0%)	**78.4**(↑3.4%)	**59.3**(↑3.8%)	**79.1**(↑1.4%)	**91.4**(↑1.9%)	**86.0**(↑1.8%)
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
Ours		**27.5**(↑5.4%)	**84.0**(↑2.1%)	**38.2**(↑7.6%)	**91.7**(↑2.8%)	**69.0**(↑2.8%)	**54.8**(↑3.4%)	**75.9**(↑4.5%)	**72.3**(↑1.5%)	**94.8**(↑0.6%)

Table 1: Cross-model comparison of hallucination mitigation methods on TruthfulQA and PDTB with DISQ evaluation. CAD and DoLa are decoding-based methods, ITI is an inference-time intervention method, and TruthX and SPACE are representation editing methods.

Metric	Qwen2-7B		Llama2-7B		DeepSeek-7B	
	H→T	T→H	H→T	T→H	H→T	T→H
T×I	3.4	5.2	2.8	3.9	3.1	6.6
MC1	3.8	4.4	3.4	3.2	9.3	10.7
MC2	1.4	2.1	4.5	5.1	2.8	7.3
True	1.9	2.7	1.5	1.7	3.9	5.9
Info	1.8	3.5	0.6	1.0	0.9	2.5

(a) Impact of bidirectional activation patching



(b) Layer-wise mitigation performance of scalar fields

Figure 4: (a) Bidirectional activation patching results on TruthfulQA. Red and green indicate improvement and degradation, respectively. H→T and T→H denote hallucination-to-truthfulness and reverse patching directions, respectively. (b) Layer-wise comparison of scalar fields for mitigation. The former is learned at each layer, while the latter is obtained through pullback.

6 Conclusion

In this paper, we propose a geometrically interpretable framework for H-Neuron identification and hallucination mitigation in LLMs. Our approach establishes a three-stage theoretical framework: (i) Based on Kantorovich–Rubinstein duality, we derive an intrinsic scalar field on the activation manifold and show that, under optimal transport assumptions, its gradient aligns with the geodesic transport direction, providing a monotonic scalar signal toward truthful activations, while decomposing the tangent vector along the activation axes yields component polarities that identify H-Neurons by revealing their pro-hallucinatory dynamical roles; (ii) leveraging this geometric correspondence, we formulate activation steering as a first-order dynamical system driven by the scalar field gradient, which moves activations toward regions with higher scalar field values; (iii) to improve scalability, we introduce a manifold pullback mechanism that recursively constructs approximate scalar fields across layers with bounded approximation error, reducing the need for independent layer-wise optimization. Experiments on TruthfulQA and PDTB demonstrate the effectiveness of the proposed framework in improving factuality and faithfulness. Neuron-level bidirectional intervention analyses further show that the identified neurons are causally related to hallucination behaviors, suggesting that the proposed scalar field formulation reveals hallucination-associated activation directions beyond conventional linear probing approaches.

AI USE STATEMENT

We disclose the use of generative AI tools in the preparation of this manuscript. The tools were used solely for light editing, language polishing, and rephrasing purposes to improve clarity, readability, and overall presentation. The research ideas, methodology, analysis, interpretations, and original content were developed entirely by the authors. Generative AI tools were not used to generate scientific claims, design experiments, conduct analyses, or replace the authors' intellectual contributions. All AI-assisted suggestions were carefully reviewed, verified, and revised by the authors to ensure accuracy, consistency, and alignment with the intended meaning of the manuscript. The authors take full responsibility for the final content and conclusions presented in this work.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. URL https://arxiv.org/abs/1610.01644.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Comput.*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746. URL https://doi.org/10.1162/089976698300017746.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Marc E. Canby, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. How reliable are causal probing interventions? In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 857–878, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. doi: 10.18653/v1/2025.ijcnlp-long.47. URL https://aclanthology.org/2025.ijcnlp-long.47/.

Chi Seng Cheang, Hou Pong Chan, Wenxuan Zhang, and Yang Deng. Do llms really know what they don't know? internal states mainly reflect knowledge recall rather than truthfulness, 2026. URL https://arxiv.org/abs/2510.09033.

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. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=Zj12nzlQbz.

Zhongzhi Chen, Xingwu Sun, Xianfeng Jiao, Fengzong Lian, Zhanhui Kang, Di Wang, and Cheng-Zhong Xu. Truth forest: Toward multi-scale truthfulness in large language models through intervention without tuning, 2024b. URL https://arxiv.org/abs/2312.17484.

Jack Choquette, Wishwesh Gandhi, Olivier Giroux, Nick Stam, and Ronny Krashinsky. Nvidia a100 tensor core gpu: Performance and innovation. *IEEE Micro*, 41(2):29–35, 2021. doi: 10.1109/MM.2021.3061394.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James

Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL http://jmlr.org/papers/v24/22-1144.html.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Th6NyL07na.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. Deepseek llm: Scaling open-source language models with longtermism, 2024. URL https://arxiv.org/abs/2401.02954.

Yongxin Deng, Zhen Fang, Sharon Li, and Ling Chen. Beyond in-domain detection: Spikescore for cross-domain hallucination detection. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Y16qXOaylp.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. Haloscope: Harnessing unlabeled llm generations for hallucination detection. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 102948–102972. Curran Associates, Inc., 2024. doi: 10.52202/079017-3270. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ba92705991cfbbcedc26e27e833ebbae-Paper-Conference.pdf.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5271–5285, 2022.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Farima Fatahi Bayat, Xin Liu, H. Jagadish, and Lu Wang. Enhanced language model truthfulness with learnable intervention and uncertainty expression. In Lun-Wei Ku, Andre Martins, and

Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12388–12400, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.737. URL https://aclanthology.org/2024.findings-acl.737/.

Cheng Gao, Huimin Chen, Chaojun Xiao, Zhiyi Chen, Zhiyuan Liu, and Maosong Sun. H-neurons: On the existence, impact, and origin of hallucination-associated neurons in llms. *arXiv preprint arXiv:2512.01797*, 2025.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, An-

nie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jiatong Han, Neil Band, Muhammed Razzak, Jannik Kossen, Tim G. J. Rudner, and Yarin Gal. Simple factuality probes detect hallucinations in long-form natural language generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16209–16226, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.880. URL https://aclanthology.org/2025.findings-emnlp.880/.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):155, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

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):138, March 2023. ISSN 1557-7341. doi: 10.1145/3571730. URL http://dx.doi.org/10.1145/3571730.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. LLM internal states reveal hallucination risk faced with a query. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 88–104, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.6. URL https://aclanthology.org/2024.blackboxnlp-1.6/.

Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. On large language models' hallucination with regard to known facts. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1041–1053, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.60. URL https://aclanthology.org/2024.naacl-long.60/.

Jushi Kai, Tianhang Zhang, Hai Hu, and Zhouhan Lin. SH2: Self-highlighted hesitation helps you decode more truthfully. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4514–4530, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.260. URL https://aclanthology.org/2024.findings-emnlp.260/.

Al Kari. Activation manifold projection: Liberating task-specific behaviors from llm architectures, 2025. URL https://arxiv.org/abs/2510.17902.

Linggang Kong, Yunlong Zhang, Xiaofeng Zhong, Haoran Fu, Yongjie Wang, and Huijun Liu. Halugnn: Hallucination detection in large language models using graph neural network. *Expert Systems with Applications*, 306:130857, 2026. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2025.130857. URL https://www.sciencedirect.com/science/article/pii/S0957417425044720.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 9332–9346, 2020.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023. URL https://arxiv.org/abs/2302.09664.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Haohang Li, Yupeng Cao, Yangyang Yu, Jordan W. Suchow, and Zining Zhu. Truth neurons, 2025a. URL https://arxiv.org/abs/2505.12182.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aLLuYpn83y.

Ruosen Li, Ziming Luo, and Xinya Du. FG-PRM: Fine-grained hallucination detection and mitigation in language model mathematical reasoning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4247–4278, Suzhou, China, November 2025b. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.228. URL https://aclanthology.org/2025.findings-emnlp.228/.

Kewei Liao, Tianbo Wang, Yuqing Ma, Zhange Zhang, Zhicheng Geng, Xiaowei Zhao, Jiakai Wang, and Xianglong Liu. Query-routed activation editing with truth-hierarchical preference optimization. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 31979–31987. AAAI Press, 2026. doi: 10.1609/AAAI.V40I38.40468. URL https://doi.org/10.1609/aaai.v40i38.40468.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Junteng Liu, Shiqi Chen, Yu Cheng, and Junxian He. On the universal truthfulness hyperplane inside LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18199–18224, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1012. URL https://aclanthology.org/2024.emnlp-main.1012/.

Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4791–4797, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.291. URL https://aclanthology.org/2023.emnlp-main.291/.

Xianxuan Long, Yao Fu, Runchao Li, Mu Sheng, Haotian Yu, Xiaotian Han, and Pan Li. When truthful representations flip under deceptive instructions?, 2025. URL https://arxiv.org/abs/2507.22149.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Wen Luo, Guangyue Peng, Wei Li, Shaohang Wei, Feifan Song, Liang Wang, Nan Yang, Xingxing Zhang, Jing Jin, Furu Wei, and Houfeng Wang. Two pathways to truthfulness: On the intrinsic encoding of llm hallucinations, 2026. URL https://arxiv.org/abs/2601.07422.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, 2020.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Yisong Miao, Hongfu Liu, Wenqiang Lei, Nancy Chen, and Min-Yen Kan. Discursive socratic questioning: Evaluating the faithfulness of language models' understanding of discourse relations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6277–6295, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.341. URL https://aclanthology.org/2024.acl-long.341/.

Mengjia Niu, Hao Li, Jie Shi, Hamed Haddadi, and Fan Mo. Mitigating hallucinations in large language models via self-refinement-enhanced knowledge retrieval, 2024. URL https://arxiv.org/abs/2405.06545.

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. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. Steer llm latents for hallucination detection, 2025. URL https://arxiv.org/abs/2503.01917.

Van-Cuong Pham and Thien Huu Nguyen. Householder pseudo-rotation: A novel approach to activation editing in llms with direction-magnitude perspective, 2024. URL https://arxiv.org/abs/2409.10053.

Angelos Poulis, Mark Crovella, and Evimaria Terzi. Testing the limits of truth directions in llms, 2026. URL https://arxiv.org/abs/2604.03754.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias (eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL https://aclanthology.org/L08-1093/.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M. Ponti, and Shay B. Cohen. Spectral editing of activations for large language model alignment, 2024. URL https://arxiv.org/abs/2405.09719.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.828. URL https://doi.org/10.18653/v1/2024.acl-long.828.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. A comprehensive survey of hallucination in large language, image, video and audio foundation models, 2024. URL https://arxiv.org/abs/2405.09589.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.69. URL https://aclanthology.org/2024.naacl-short.69/.

Samuel Soo, Chen Guang, Wesley Teng, Chandrasekaran Balaganesh, Tan Guoxian, and Yan Ming. Interpretable steering of large language models with feature guided activation additions, 2025. URL https://arxiv.org/abs/2501.09929.

Federico Tiblias, Irina Bigoulaeva, Jingcheng Niu, Simone Balloccu, and Iryna Gurevych. Hypothesis-driven feature manifold analysis in llms via supervised multi-dimensional scaling, 2026. URL https://arxiv.org/abs/2510.01025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023a. URL https://arxiv.org/abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023b. URL https://arxiv.org/abs/2307.09288.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias, 2020. URL https://arxiv.org/abs/2004.12265.

Cédric Villani. *Optimal transport – Old and new*, volume 338, pp. xxii+973. 01 2008. doi: 10.1007/978-3-540-71050-9.

Pengbo Wang, Chaozhuo Li, Chenxu Wang, Liwen Zheng, Litian Zhang, and Xi Zhang. One space to rule them all: Jointly mitigating factuality and faithfulness hallucinations in llms, 2025a. URL https://arxiv.org/abs/2506.11088.

Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the ACM on Web Conference 2025*, WWW 25, pp. 25622578. ACM, April 2025b. doi: 10.1145/3696410.3714640. URL http://dx.doi.org/10.1145/3696410.3714640.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for llms via dynamic steering vectors, 2025c. URL https://arxiv.org/abs/2410.12299.

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.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models, 2024. URL https://arxiv.org/abs/2308.03958.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Jingyuan Yang, Rongjun Li, Weixuan Wang, Ziyu Zhou, Zhiyong Feng, and Wei Peng. Lf-steering: Latent feature activation steering for enhancing semantic consistency in large language models, 2025. URL https://arxiv.org/abs/2501.11036.

Zeping Yu and Sophia Ananiadou. Neuron-level knowledge attribution in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3267–3280, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.191. URL https://aclanthology.org/2024.emnlp-main.191/.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods, 2024. URL https://arxiv.org/abs/2309.16042.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How language model hallucinations can snowball, 2023. URL https://arxiv.org/abs/2305.13534.

Shaolei Zhang, Tian Yu, and Yang Feng. TruthX: Alleviating hallucinations by editing large language models in truthful space. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8908–8949, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.483. URL https://aclanthology.org/2024.acl-long.483/.

Andy Zou, Long Phan, Sarah Li Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *CoRR*, abs/2310.01405, 2023. doi: 10.48550/ARXIV.2310.01405. URL https://doi.org/10.48550/arXiv.2310.01405.

Andy Zou, Long Phan, Sarah Li Chen, James Campbell, Phillip Huang Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Annah Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Troy Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J Zico Kolter, and Dan Hendrycks. Enhancing neural network transparency through representation analysis, 2024. URL https://openreview.net/forum?id=aCgybhcZFi.

APPENDIX CONTENTS

A Related work **20**

A.1 Hallucination Detection 20

A.2 Hallucination Mitigation 21

B Additional Background and Proofs of Lemmas **21**

B.1 Fisher Information Metric on Activation Manifolds. 22

B.2 Geometric Objects and Tensor Notation 23

B.3 Proof of Lemma 1 27

B.4 Proof of Lemma 2 28

B.5 Proof of Lemma 3 29

C Kantorovich Dual Construction and Properties of the Scalar Field **30**

C.1 Kantorovich-Rubinstein Duality and Saturation Condition 30

C.2 Proof of Theorem: Equivalence Between Geodesics and Scalar Field Tangent Vectors 32

C.3 Proof of Corollary: Scalar Field Based Identification of H-Neurons 34

C.4 Proof of Corollary: Monotonicity of the Scalar Field along Geodesics 35

C.5 Robustness of the Scalar Field 37

D Proof of Pullback of Scalar Fields Across Layers **39**

D.1 Proof of Proposition: Recursive Construction via Normalized Pullback 39

D.2 Proof of Proposition: Cross-Layer Structural Suboptimality Bound 40

E Experimental Setup **43**

E.1 Datasets and Metrics. 43

E.2 Baselines 43

E.3 Implementation Details. 43

E.4 Practical Training of the Scalar Field. 44

E.5 Practical ODE Intervention. 45

F Experimental Validation **45**

F.1 Cross-Model Comparison with Baselines 45
F.2 Ablation Experiment 47
F.3 Effect of Intervention Strength 48
F.4 Computational Cost of Pullback versus Scalar Field Optimization 48
F.5 Comparison with Existing Activation Steering Baselines 49
F.6 Geometric Visualization of the Scalar Field 49
F.7 Consistency between Probe and Scalar field Neuron Identification 49
F.8 Saturation Behavior under Increasing Sampling Size 50
F.9 Comparison of Different Distance Metrics 50
F.10 Trajectory Consistency of Scalar Field-Guided Mitigation 50
F.11 Empirical Verification of Lipschitz Constants 51
F.12 Hyperparameter Experimentation 51

G Limitations **52**

A Related work

A.1 Hallucination Detection

Early research on LLM hallucinations primarily focused on viewing them as a purely output-level phenomenon Lin et al. (2022); Ji et al. (2023); benchmarks and surveys such as TruthfulQA and the work by Ji et al. systematized this perspective Huang et al. (2025); Sahoo et al. (2024); Zhang et al. (2023). These studies emphasized factual inconsistencies and fabricated content, rather than the underlying interpretable internal mechanisms Maynez et al. (2020); Kryściński et al. (2020); Dziri et al. (2022); Li et al. (2025b); Jiang et al. (2024). More recently, some researchers have begun to look beyond merely evaluating output content, demonstrating that hallucinations already exist within structured internal representations and can be predicted by probing the hidden states during the generation process Ji et al. (2024); Alain & Bengio (2018); Du et al. (2024); Kong et al. (2026). Building upon this foundation, studies employing methods such as "Inference-Time Intervention" and "The Geometry of Truth" have revealed that these internal states possess separable geometric structures Long et al. (2025); Luo et al. (2026); Park et al. (2025); shifting activations along specific linear directions within the latent space can significantly enhance truthfulness Li et al. (2023); Marks & Tegmark (2024). These findings support the broader view that high-level semantic attributes can be captured as specific directions within the representation space Turner et al. (2024); Rimsky et al. (2024); Zou et al. (2023); Poulis et al. (2026). These findings at the representational level naturally lead to a more fine-grained mechanistic question: which specific internal units align with these "truthful" or "hallucinatory" directions? Li et al. (2025a); Yu & Ananiadou (2024) The literature on knowledge localization establishes a theoretical prior that semantic behaviors are concentrated within sparse neural subsets Dai et al. (2022). Guided by this insight, recent studies utilizing linear probing have identified "hallucination-related neurons" (H-Neurons) Gao et al. (2025). These neurons constitute an extremely sparse subset, accounting for less than 0.1% of the total units within the feed-forward networks. Furthermore, they bear a causal relationship to hallucinations: artificially boosting their activation exacerbates over-compliance, whereas suppressing them mitigates this risk Gao et al. (2025). Despite these advancements, existing detection methods still face three fundamental limitations Cheang et al. (2026). First, they rely on sparse linear probes, which primarily serve as statistical discriminators and offer limited mechanistic interpretability Alain & Bengio (2018). Second, these probing methods exhibit poor generalization across different activation spaces and layers, typically necessitating repeated training Gao et al. (2025); Zou et al. (2023).

A.2 HALLUCINATION MITIGATION

While the above studies have shifted hallucination research from output-level evaluation to internal representation Liu et al. (2023) and neuron-level localization, detection alone is insufficient Zou et al. (2023); Gao et al. (2025); Liu et al. (2024); Han et al. (2025). A more important goal is hallucination mitigation: once hallucination-related states or units are identified, how can the model be guided back from distorted representations to truthful ones? Turner et al. (2024); Rimsky et al. (2024); Wang et al. (2025b); Yang et al. (2025) Early mitigation methods mainly operated in the input space, such as context-constrained prompt engineering Wei et al. (2022) and Retrieval-Augmented Generation (RAG) Lewis et al. (2020); Niu et al. (2024). Although often effective, these approaches constrain the model externally and do not directly address the internal mechanisms underlying hallucination. With advances in mechanistic interpretability, research has gradually shifted toward white-box intervention on internal model states Soo et al. (2025). This shift is supported by an emerging understanding of the macro-structure of representation space. Probing studies Li et al. (2023) show that hidden states associated with truthful and hallucinated content are often geometrically separable by a hyperplane in latent space. This hyperplane defines not only the boundary between truth and falsehood, but also a normal direction that can be viewed as a semantic axis from hallucination to truth. Accordingly, hallucination mitigation can be understood as steering internal representations along meaningful latent directions rather than merely correcting outputs after generation. Based on this idea, methods such as Activation Addition Turner et al. (2024) perform preliminary intervention by injecting concept vectors along such directions during inference. To achieve finer control, subsequent studies have moved from macroscopic geometric directions to microscopic functional units Zhang & Nanda (2024). Recent evidence suggests that hallucinated outputs are causally linked to specific neuronal subsets: under misleading or leading contexts, these neurons exhibit excessive sycophantic behavior Sharma et al. (2024); Wei et al. (2024), producing abnormal activations that push the model away from factual consistency. This sycophancy-driven causal mechanism motivates a more targeted mitigation strategy, namely, locating and suppressing hallucination-related neurons during inference to directly recalibrate the models internal state. Despite this progress, current hallucination mitigation methods still have notable limitations. Detection and intervention remain theoretically disconnected: existing methods may identify separable directions or salient neurons, but they do not provide a unified account of how hallucinated representations should be transformed into truthful ones. Moreover, many mitigation strategies remain heuristic, relying on empirically discovered linear directions or manually designed neuron-level suppression rules Li et al. (2023); Marks & Tegmark (2024). Consequently, they still lack a principled mechanistic framework for characterizing the optimal correction direction and its propagation across layers, neurons, and representation spaces.

B ADDITIONAL BACKGROUND AND PROOFS OF LEMMAS

Definition 2 (**Recursive Cross-Layer Mapping**). *Take the mapping from the final activation state of Transformer layer $L-1$ to that of Transformer layer L as an example. Let $H_{L-1} \in \mathcal{M}_{L-1}$ and R_{L-1} respectively denote the activation state and the residual stream state at the $(L-1)$-th layer. The recursive cross-layer mapping $\phi_{L-1} : \mathcal{M}_{L-1} \to \mathcal{M}_L$ is then defined as*

- *The residual stream for the subsequent layer is updated as $R_L = R_{L-1} + \mathrm{LinearOut}_L(H_{L-1}) + \mathrm{Attn}_L(\mathrm{LN}(R_{L-1} + \mathrm{LinearOut}_L(H_{L-1})))$, and the cross-layer mapping $\phi_{L-1}(H_{L-1}) = H_L$ is then defined by projecting this updated residual stream into the activation space of the L-th layer: $H_L = \sigma(\mathrm{LinearIn}_L(\mathrm{LN}(R_L)))$.*

Definition 3 (**Hallucination-Associated Neurons (H-Neurons)**). *Hallucination Associated Neurons (H-Neurons) constitute a sparse neuronal subset whose functional dynamics systematically degrade output faithfulness in favor of hallucinatory generation. We do not claim that H-Neurons correspond to anatomically isolated hallucination circuits. Instead, they represent activation dimensions whose perturbations consistently influence hallucination-related behavior. Given the hidden activation state at layer L, parameterized as*

$$H^{(L)} = (x^1, \ldots, x^\mu, \ldots, x^d) \in \mathcal{M}_L, \tag{15}$$

where x^μ denotes the activation scalar of the μ-th neuron. An H-Neuron is characterized not merely by its activation amplitude, but by a discriminative structural contribution that mathematically isolates hallucinatory trajectories from faithful inference.

To explicitly formalize this mechanism, let $\mathbf{w}_\mu^{\text{down}} \in \mathbb{R}^d$ denote its corresponding down projection vector. The isolated projection of this neuron onto the residual stream is quantified as:

$$\mathbf{h}^{(\mu)} = x^\mu \mathbf{w}_\mu^{\text{down}}. \tag{16}$$

Accordingly, its normalized functional contribution $\mathcal{C}_\mu$ with respect to the aggregated hidden representation $\mathbf{h}$ is defined via the L_2 norm ratio:

$$\mathcal{C}_\mu = \frac{\left\|\mathbf{h}^{(\mu)}\right\|_2}{\|\mathbf{h}\|_2} = \frac{|x^\mu| \left\|\mathbf{w}_\mu^{\text{down}}\right\|_2}{\|\mathbf{h}\|_2}. \tag{17}$$

This formulation establishes that the functional contribution of a neuron is directly governed by its dynamic activation state.

Consequently, the hallucination associated identity of a neuron is deterministically defined by its causal behavioral response under activation intervention. Applying an affine scaling scalar $\alpha \in \mathbb{R}^+$ to the activation, $x^\mu \leftarrow \alpha x^\mu$, induces a proportionally perturbed partial representation $\mathbf{h}^{(\mu)}(\alpha) = \alpha \mathbf{h}^{(\mu)}$. The resulting normalized contribution becomes:

$$\mathcal{C}_\mu(\alpha) = \frac{\alpha \left\|\mathbf{h}^{(\mu)}\right\|_2}{\|\mathbf{h}(\alpha)\|_2}. \tag{18}$$

Under the asymptotic assumption that the isolated neuronal contribution remains marginal relative to the global residual stream, the denominator shift is bounded, yielding $\|\mathbf{h}(\alpha)\|_2 \approx \|\mathbf{h}\|_2$. Thus, the perturbed contribution converges to a linear approximation:

$$\mathcal{C}_\mu(\alpha) \approx \alpha \mathcal{C}_\mu. \tag{19}$$

A neuron is strictly classified as an H-Neuron if directly manipulating its activation value causally dictates the hallucinatory propensity of the model. Specifically, targeted amplification of the activation scalar ($\alpha > 1$) exacerbates over-compliant or unfaithful generation, whereas targeted suppression of the activation ($\alpha < 1$) deterministically guides the generative state toward rigorous and factually grounded inference.

B.1 Fisher Information Metric on Activation Manifolds.

To instantiate the Riemannian metric on the activation manifold, we adopt the Fisher information metric induced by the output distribution of the frozen LLM. This metric provides a geometry informed by the prediction task and measures the sensitivity of model outputs to perturbations of the internal representation. We do not claim that it is the unique intrinsic metric of LLM representations.

Consistent with the empirical measures introduced above, $H_L = (x^1, \ldots, x^d) \in \mathcal{M}_L$ denotes the activation state of a response token at the L-th layer, corresponding to $h_{L,t}^{(s)}$. Given H_L, the subsequent frozen layers define a conditional distribution over the next token. Throughout this section, z denotes the next token. For simplicity, the prompt and preceding tokens are omitted from the notation $p_\Theta(z \mid H_L)$, while all other model states are kept fixed.

The Fisher metric tensor is defined as

$$g_{\mu\nu}(H_L) = \mathbb{E}_{z \sim p_\Theta(\cdot \mid H_L)} \left[\frac{\partial \log p_\Theta(z \mid H_L)}{\partial x^\mu} \frac{\partial \log p_\Theta(z \mid H_L)}{\partial x^\nu} \right]. \tag{20}$$

Here, x^μ denotes the μ-th activation coordinate. The Fisher metric describes the local change in the output distribution induced by perturbations of the activation state. Our method therefore requires only the metric in the neighborhood of activation states used during training and intervention, rather than an explicit construction of the entire statistical manifold. This follows the standard Fisher–Rao formulation, in which the metric is determined by the second moment of the score Amari (1998).

For the discrete vocabulary space $\mathcal{V}$, define the local score component as

$$s_\mu(z; H_L) = \frac{\partial \log p_\Theta(z \mid H_L)}{\partial x^\mu}. \tag{21}$$

The Fisher metric can then be written as

$$g_{\mu\nu}(H_L) = \sum_{z\in\mathcal{V}} p_\Theta(z \mid H_L) s_\mu(z; H_L) s_\nu(z; H_L). \tag{22}$$

Because exact summation over the vocabulary is computationally expensive, we draw M independent samples $z_m \sim p_\Theta(\cdot \mid H_L)$ and use the following Monte Carlo estimate:

$$g_{\mu\nu}(H_L) \approx \frac{1}{M} \sum_{m=1}^{M} s_\mu(z_m; H_L) s_\nu(z_m; H_L) + \xi I_{\mu\nu}, \qquad \xi > 0. \tag{23}$$

Here, $I_{\mu\nu}$ denotes the identity matrix in the adopted activation coordinates. The damping term $\xi I_{\mu\nu}$ ensures that the estimated metric is positive definite even when the Monte Carlo approximation is rank deficient. Since the estimated Fisher metric is composed of a low-rank data-dependent term and a positive diagonal damping term, we write

$$g(H_L) = S(H_L) S(H_L)^\top + \xi I, \tag{24}$$

where

$$S(H_L) = \frac{1}{\sqrt{M}} \left[s(z_1; H_L), \ldots, s(z_M; H_L) \right].$$

The inverse metric tensor $g^{\mu\nu}(H_L)$ is defined by $g^{\mu\alpha}(H_L) g_{\alpha\nu}(H_L) = \delta^\mu_{\ \nu}$.

For the scalar field f^*_L, the tangent vector induced by the Riemannian metric is given by

$$v^\mu(H_L) = g^{\mu\nu}(H_L) \nabla_\nu f^*_L(H_L) = \left(g(H_L)^{-1} \nabla f^*_L(H_L) \right)^\mu. \tag{25}$$

Directly constructing and inverting the full $d \times d$ metric is computationally prohibitive. Instead, using the Woodbury identity, the metric-adjusted scalar field tangent vector can be efficiently computed as

$$v(H_L) = \xi^{-1} \nabla f^*_L(H_L) - \xi^{-1} S(H_L) \left(I_M + \xi^{-1} S(H_L)^\top S(H_L) \right)^{-1} S(H_L)^\top \xi^{-1} \nabla f^*_L(H_L). \tag{26}$$

This formulation avoids explicitly forming the inverse metric tensor $g^{\mu\nu}$. Instead, it computes its action on the scalar field gradient directly.

B.2 Geometric Objects and Tensor Notation

This section provides the basic geometric notation used in the main paper. We focus on the activation manifold $\mathcal{M}_L$ of the L-th layer, where each attainable hidden activation state is treated as a point on a smooth d-dimensional manifold. For any local chart on $\mathcal{M}_L$, we denote the coordinate representation of an activation state by $(x^1, \ldots, x^\mu, \ldots, x^d)$, where x^μ corresponds to the activation value of the μ-th neuron. Throughout this paper, Latin indices such as a, b, c are used as abstract tensor indices, while Greek indices such as μ, ν, ρ are used for coordinate components associated with neuron activation axes.

B.2.1 Tangent vectors and tangent spaces.

For a point $p \in \mathcal{M}_L$, a tangent vector at p is intrinsically viewed as a derivation acting on smooth scalar functions defined near p. Specifically, a tangent vector v_p is a linear map on smooth function germs satisfying the Leibniz rule:

$$v_p(fh) = v_p(f) h(p) + f(p) v_p(h). \tag{27}$$

Thus, $v_p(f)$ represents the directional derivative of the function f at the activation state p. The collection of all tangent vectors at p forms the tangent space $T_p\mathcal{M}_L$. In local activation coordinates $(x^1, \ldots, x^d)$, the natural coordinate-induced basis of $T_p\mathcal{M}_L$ is given by $\{(\partial/\partial x^\mu)^a|_p\}_{\mu=1}^d$, and any tangent vector $v^a \in T_p\mathcal{M}_L$ can be expanded as:

$$v^a = v^\mu \left(\frac{\partial}{\partial x^\mu}\right)^a. \tag{28}$$

Here, v^μ denotes the component of v^a along the μ-th activation axis.

To explicitly compute the components v^ν of a given tangent vector v^a, we rely on its intrinsic action as a derivation. In a coordinate-free perspective, applying a tangent vector to a scalar function corresponds to contracting the vector with the differential (or gradient) of that function. To extract a specific component, we consider the vector's action on the ν-th local coordinate function x^ν. By contracting v^a with the corresponding natural basis 1-form $(dx^\nu)_a$ of the cotangent space $T_p^*\mathcal{M}_L$, and substituting the expansion from Eq. (28), we obtain:

$$v^a(dx^\nu)_a = \left[\sum_{\mu=1}^d v^\mu \left(\frac{\partial}{\partial x^\mu}\right)^a\right](dx^\nu)_a = \sum_{\mu=1}^d v^\mu \left[\left(\frac{\partial}{\partial x^\mu}\right)^a (dx^\nu)_a\right] = \sum_{\mu=1}^d v^\mu \delta^\nu_\mu = v^\nu. \tag{29}$$

Equivalently, viewed as a differential operator acting directly on the coordinate function x^ν without explicit index contraction, this relation is straightforwardly expressed as $v_p(x^\nu) = \sum_{\mu=1}^d v^\mu(\partial x^\nu/\partial x^\mu) = v^\nu$. This formal derivation clarifies exactly how the geometric object v^a projects its values onto the local activation axes.

The disjoint union of all tangent spaces defines the tangent bundle: $T\mathcal{M}_L = \bigsqcup_{p\in\mathcal{M}_L} T_p\mathcal{M}_L$. A smooth vector field is a section of this bundle, assigning to each activation state p a tangent vector $v^a|_p \in T_p\mathcal{M}_L$.

B.2.2 Cotangent vectors and one-forms.

The cotangent space at p, denoted by $T_p^*\mathcal{M}_L$, is the dual space of the tangent space $T_p\mathcal{M}_L$. It is defined as the set of all linear functionals on $T_p\mathcal{M}_L$:

$$T_p^*\mathcal{M}_L = \mathrm{Hom}(T_p\mathcal{M}_L, \mathbb{R}). \tag{30}$$

An element $\omega_a \in T_p^*\mathcal{M}_L$ is called a cotangent vector (or covector). Its linear action on a tangent vector v^a returns a scalar, denoted by the contraction $\omega_a v^a \in \mathbb{R}$. In local activation coordinates, the dual basis $\{(dx^\mu)_a\}_{\mu=1}^d$ is defined to be dual to the natural tangent basis (Eq. 28) by the relation:

$$(dx^\mu)_a \left(\frac{\partial}{\partial x^\nu}\right)^a = \delta^\mu_\nu, \tag{31}$$

where δ^μ_ν is the Kronecker delta. Any covector can be written as $\omega_a = \sum_{\mu=1}^d \omega_\mu (dx^\mu)_a$. Analogous to tangent vectors, the specific component ω_ν of a covector can be explicitly extracted by evaluating it on the natural basis vectors of the tangent space. Using the dual relation, we have:

$$\omega_a \left(\frac{\partial}{\partial x^\nu}\right)^a = \left[\sum_{\mu=1}^d \omega_\mu (dx^\mu)_a\right]\left(\frac{\partial}{\partial x^\nu}\right)^a = \sum_{\mu=1}^d \omega_\mu \left[(dx^\mu)_a \left(\frac{\partial}{\partial x^\nu}\right)^a\right] = \sum_{\mu=1}^d \omega_\mu \delta^\mu_\nu = \omega_\nu. \tag{32}$$

For a smooth scalar field $f : \mathcal{M}_L \to \mathbb{R}$, its covariant derivative (or gradient) $\nabla_a f$ is a smooth cotangent vector field. Since f is a scalar, the covariant derivative reduces to the ordinary partial derivative in local coordinates, i.e., $\nabla_\mu f = \partial f/\partial x^\mu$. Applying the component extraction principle to this gradient, the component $\nabla_\mu f$ is obtained by contracting the one-form $\nabla_a f$ with the μ-th coordinate basis vector. Expanding $\nabla_a f = \sum_{\nu=1}^d (\nabla_\nu f)(dx^\nu)_a$, this projection is precisely given by:

$$\left(\frac{\partial}{\partial x^\mu}\right)^a \nabla_a f = \sum_{\nu=1}^{d} (\nabla_\nu f) \left[(dx^\nu)_a \left(\frac{\partial}{\partial x^\mu}\right)^a \right] = \sum_{\nu=1}^{d} (\nabla_\nu f) \delta^\nu_\mu = \nabla_\mu f. \tag{33}$$

This demonstrates how the geometric object $\nabla_a f$ translates into the standard partial derivatives along the local activation axes.

B.2.3 Tensor fields and algebraic operations.

More generally, a tensor of type (k, l) at a point $p \in \mathcal{M}_L$ is defined as a multilinear map acting on the product of k cotangent spaces and l tangent spaces. The set of all such tensors forms the space:

$$T_p^{(k,l)}\mathcal{M}_L = \underbrace{T_p\mathcal{M}_L \otimes \cdots \otimes T_p\mathcal{M}_L}_{k} \otimes \underbrace{T_p^*\mathcal{M}_L \otimes \cdots \otimes T_p^*\mathcal{M}_L}_{l}. \tag{34}$$

In abstract index notation, a tensor of this type is denoted by $T^{a_1 \cdots a_k}{}_{b_1 \ldots b_l}$. In local coordinates $\{x^\mu\}$, generalizing Eq. 28 and Eq. 31, it admits a unique expansion in terms of the tensor product basis:

$$T = T^{\mu_1 \cdots \mu_k}{}_{\nu_1 \ldots \nu_l} \frac{\partial}{\partial x^{\mu_1}} \otimes \cdots \otimes \frac{\partial}{\partial x^{\mu_k}} \otimes dx^{\nu_1} \otimes \cdots \otimes dx^{\nu_l}. \tag{35}$$

The coefficients $T^{\mu_1 \cdots \mu_k}{}_{\nu_1 \ldots \nu_l}$ are smooth functions representing the tensor components. Under this formalism, a scalar field is a $(0, 0)$-tensor, a vector field is a $(1, 0)$-tensor, and a one-form is a $(0, 1)$-tensor.

Metric and Contraction. The Riemannian metric g is a symmetric, non-degenerate $(0, 2)$-tensor field, denoted by g_{ab} in abstract indices. The primary algebraic operation connecting different tensor types is **contraction**. A contraction is a trace-like operation that maps a (k, l) tensor to a $(k-1, l-1)$ tensor by evaluating a covariant slot on a contravariant slot.

Abstract index notation makes this operation highly intuitive: repeating an upper and a lower index implies contraction over those specific slots. For instance, contracting the second contravariant and first covariant slots of a $(2, 1)$-tensor $T^{ab}{}_c$ yields a $(1, 0)$-tensor (a vector), simply denoted as $S^a = T^{ab}{}_b$.

A critical instance of contraction in our geometric framework is the inner product. Given two tangent vectors u^a and v^a, the metric g_{ab} induces a scalar via double contraction:

$$\langle u, v \rangle_g = g_{ab} u^a v^b. \tag{36}$$

In local component formulation, this is evaluated as $g_{\mu\nu} u^\mu v^\nu$.

Index Raising and Lowering. The non-degeneracy of g_{ab} implies the existence of an inverse metric tensor g^{ab} of type $(2, 0)$, defined by the contraction $g_{ac} g^{cb} = \delta^b_a$. These tensors define the musical isomorphisms between $T_p\mathcal{M}_L$ and $T_p^*\mathcal{M}_L$:

- **Lowering:** For a vector v^a, the corresponding covector is defined by contracting with the metric: $v_a = g_{ab} v^b$.
- **Raising:** For a covector ω_a, the corresponding vector is defined by contracting with the inverse metric: $\omega^a = g^{ab} \omega_b$.

B.2.4 Abstract index notation.

We use abstract index notation (Penrose notation) to express coordinate-free tensor relations. In this formalism, Latin indices (e.g., a, b, c) do not denote numerical coordinate components. Instead, they specify tensorial slots, variance, and contraction patterns. For example, v^a denotes a tangent vector, ω_a denotes a cotangent vector, and $T^a{}_b$ denotes a tensor of type $(1, 1)$. The index a is not a coordinate label; it indicates the presence of a contravariant or covariant tensor slot.

The algebraic manipulation of abstract indices follows strict rules:

Free Indices and Tensor Type. Free indices (those appearing exactly once in a term) determine the tensor type of an expression and must match identically on both sides of any valid tensorial equation. For example, $A^a = B^a + C^a$ is well-typed because all terms are $(1,0)$-tensors. In contrast, $A^a = B_a + C^a$ is ill-defined unless a metric is explicitly invoked to convert the covector B_a into a vector.

Contraction Rule. A repeated upper-lower pair of the identical index denotes contraction over that tensorial slot. The contracted index becomes a dummy index and "disappears" from the resulting tensor type. For example:

- $\omega_a v^a$ denotes the natural pairing between a covector and a vector, yielding a scalar (type $(0,0)$).
- $T^a{}_b v^b = w^a$ represents a $(1,1)$-tensor acting on a vector. The index b is contracted, leaving the free index a, yielding a $(1,0)$-tensor.
- $A^a{}_b B^b{}_c = C^a{}_c$ is a composition where b is contracted, resulting in a $(1,1)$-tensor.

Dummy Index Substitution. A dummy (contracted) index can be systematically renamed without changing the expression's meaning. Thus, $\omega_a v^a = \omega_b v^b$ represents the identical scalar quantity. However, free indices cannot be renamed independently; renaming a free index must be done simultaneously across all terms in the equation (e.g., $T^a{}_b v^b = w^a \iff T^c{}_b v^b = w^c$).

The "At Most Twice" Rule. Within a single monomial term, an index should appear at most twice: exactly once as an upper index and exactly once as a lower index. Expressions like $T^a{}_a$ validly denote a trace (contraction). However, expressions like $T^a{}_a v^a$ are ambiguous and strictly forbidden; they must be rewritten with distinct dummy indices, such as $(T^b{}_b)v^a$.

Tensor Products vs. Contractions. If tensors are multiplied without repeating indices, the result is their outer tensor product. For instance, $v^a \omega_b$ is a tensor of type $(1,1)$. Repeating an index induces a contraction: $v^a \omega_a$ is a scalar.

Metric Isomorphisms and Covariant Differentiation. The metric tensor provides the canonical mechanism for index manipulation without reference to coordinates. Lowering is written as $v_a = g_{ab} v^b$, and raising as $\omega^a = g^{ab} \omega_b$. Furthermore, the covariant derivative operator ∇_a structurally adds one covariant index to the object on which it acts. For a scalar field f, $\nabla_a f$ is a $(0,1)$-tensor. In this sense, the derivative index a creates a new covariant slot.

Combining these rules, the tangent vector used in the main paper is expressed abstractly as:

$$v^a = g^{ab} \nabla_b f_L^*. \tag{37}$$

This establishes the local mitigation direction as a geometrically consistent entity on the activation manifold, entirely independent of how neuron activations are parameterized.

B.2.5 Component notation.

Component notation is obtained once a specific local coordinate chart is chosen. We use Greek indices (e.g., μ, ν) to denote the actual numerical components along the chosen coordinate axes (the neuron activation dimensions).

While abstract indices represent the object intrinsically, component indices describe its representation in a basis. A vector field is explicitly written as $v = v^\mu \frac{\partial}{\partial x^\mu}$, where v^μ are the scalar component functions. A one-form is $\omega = \omega_\mu dx^\mu$, and a $(1,1)$-tensor is $T = T^\mu{}_\nu \frac{\partial}{\partial x^\mu} \otimes dx^\nu$.

This distinction clarifies our activation-manifold formulation. The abstract geometric equation (Eq. 37) guarantees that the mitigation flow is intrinsically well-defined. Once projected onto the local activation coordinates $\{x^\mu\}$, it yields the explicit, computable component formula:

$$v^\mu = g^{\mu\nu} \frac{\partial f_L^*}{\partial x^\nu}. \tag{38}$$

The abstract expression secures mathematical rigor, while the component expression dictates the actual neural computation.

To reinforce this difference, consider coordinate transformations $x^\mu \to x'^\mu$. While the abstract vector v^a remains invariant, its components transform according to the Jacobian matrices. A vector transforms contravariantly:

$$v'^\mu = \frac{\partial x'^\mu}{\partial x^\alpha} v^\alpha, \tag{39}$$

whereas a covector transforms covariantly:

$$\omega'_\mu = \frac{\partial x^\alpha}{\partial x'^\mu} \omega_\alpha. \tag{40}$$

More generally, a $(2,1)$-tensor transforms slot-by-slot:

$$T'^{\mu\nu}{}_\sigma = \frac{\partial x'^\mu}{\partial x^\alpha} \frac{\partial x'^\nu}{\partial x^\beta} \frac{\partial x^\gamma}{\partial x'^\sigma} T^{\alpha\beta}{}_\gamma. \tag{41}$$

In summary, abstract index notation describes the intrinsic algebra of tensor fields on the activation manifold, while component notation dictates their explicit coordinate realization along neuron activation axes. We utilize Latin indices to state fundamental geometric relationships (e.g., Eq. 37) and Greek indices to derive the actionable neuron-wise metrics (e.g., Eq. 38) required for H-Neuron detection and mitigation.

B.3 Proof of Lemma 1

Lemma 1 (**Tensor Covariant Derivative and Connection Coefficients**). *Let $\mathcal{M}$ be a Riemannian manifold equipped with a connection ∇ and Christoffel symbols Γ^k_{ij}. For any general tensor field T of type (k,l), the covariant derivative $\nabla_a : \mathcal{F}_\mathcal{M}(k,l) \to \mathcal{F}_\mathcal{M}(k,l+1)$ is defined by:*

$$\nabla_a T^{b_1 \dots b_k}{}_{c_1 \dots c_l} = \partial_a T^{b_1 \dots b_k}{}_{c_1 \dots c_l} + \sum_{i=1}^{k} \Gamma^{b_i}_{ao} T^{b_1 \dots o \dots b_k}{}_{c_1 \dots c_l} - \sum_{j=1}^{l} \Gamma^{o}_{ac_j} T^{b_1 \dots b_k}{}_{c_1 \dots o \dots c_l}. \tag{42}$$

Specifically, for a scalar field $f \in \mathcal{F}_\mathcal{M}(0,0)$, the connection terms vanish, reducing the covariant derivative to the partial derivative $\nabla_a f = \partial_a f$. In the context of neural activations x^μ, this identity implies $\nabla_\mu f = \partial f / \partial x^\mu$.

Proof. Consider a local coordinate system $(x^1, \dots, x^d)$ on $\mathcal{M}$ with the corresponding coordinate basis $\{e_b = \partial/\partial x^b\}$ and its dual basis $\{\theta^c = dx^c\}$. By the definition of the connection coefficients, the action of the covariant derivative on the basis vectors is given by $\nabla_a e_b = \Gamma^d_{ab} e_d$. To derive the action on the dual basis, we utilize the orthogonality relation $\theta^c(e_b) = \delta^c_b$. Applying the Leibniz rule to the covariant derivative of this identity, we have:

$$0 = \nabla_a \left(\theta^c(e_b)\right) = (\nabla_a \theta^c)(e_b) + \theta^c(\nabla_a e_b). \tag{43}$$

Substituting the basis transformation $\nabla_a e_b = \Gamma^d_{ab} e_d$ into the above equation yields $(\nabla_a \theta^c)(e_b) = -\Gamma^c_{ab}$, which implies the dual transformation rule:

$$\nabla_a \theta^c = -\Gamma^c_{ad} \theta^d. \tag{44}$$

A general tensor field $T \in \mathcal{F}_\mathcal{M}(k,l)$ can be expanded in terms of these bases as $T = T^{b_1 \dots b_k}{}_{c_1 \dots c_l} e_{b_1} \otimes \cdots \otimes \theta^{c_l}$. The covariant derivative $\nabla_a T$ is evaluated by applying the Leibniz rule across the tensor product. This process generates three distinct contributions: (i) the partial derivative of the component function $\partial_a T^{b_1 \cdots}_{\cdots c_l}$, (ii) k positive connection terms arising from the contravariant vectors $\{e_{b_i}\}$, and (iii) l negative connection terms arising from the covariant covectors $\{\theta^{c_j}\}$.

Summing these contributions yields the explicit expression:

$$\nabla_a T^{b_1 \dots b_k}{}_{c_1 \dots c_l} = \partial_a T^{b_1 \dots b_k}{}_{c_1 \dots c_l} + \sum_{i=1}^{k} \Gamma^{b_i}_{ao} T^{b_1 \dots o \dots b_k}{}_{c_1 \dots c_l} - \sum_{j=1}^{l} \Gamma^{o}_{ac_j} T^{b_1 \dots b_k}{}_{c_1 \dots o \dots c_l}. \tag{45}$$

For the zero-rank case (scalar field f), the absence of indices precludes any connection corrections. Thus, $\nabla_a f = \partial_a f$, which in activation space coordinates simplifies to $\nabla_\mu f = \partial_\mu f = \frac{\partial f}{\partial x^\mu}$, completing the proof. □

B.4 PROOF OF LEMMA 2

Lemma 2 (Non-degeneracy and Index Raising). *For any Riemannian manifold $(\mathcal{M}_L, g_{ab})$, the metric tensor is non-degenerate, so there exists a unique inverse metric g^{ab} satisfying $g_{ac}g^{cb} = \delta_a{}^b$, where $\delta_a{}^b$ denotes the Kronecker delta. For any smooth scalar field $f \in C^\infty(\mathcal{M}_L)$, the covector field $\nabla_b f$ can be uniquely raised to a tangent vector field $v^a := g^{ab}\nabla_b f$. Given activation state $\{x^\mu\} = (x^1, \ldots, x^\mu, \ldots, x^d)$. this vector field can be written as:*

$$v^a = v^\mu \left(\frac{\partial}{\partial x^\mu}\right)^a, \qquad v^\mu = (dx^\mu)_a v^a = g^{\mu\nu}\nabla_\nu f. \tag{46}$$

Proof. Since $(\mathcal{M}_L, g_{ab})$ is a Riemannian manifold, for each point $p \in \mathcal{M}_L$, the metric tensor $g_{ab}|_p$ defines a symmetric positive-definite bilinear form on the tangent space $T_p\mathcal{M}_L$. Positive-definiteness implies non-degeneracy. That is, if a tangent vector $u^a \in T_p\mathcal{M}_L$ satisfies

$$g_{ab}u^a w^b = 0, \qquad \forall w^b \in T_p\mathcal{M}_L, \tag{47}$$

then $u^a = 0$. Therefore, in any local coordinate chart, the metric component matrix $(g_{\mu\nu})$ is invertible. Hence, there exists a unique inverse matrix $(g^{\mu\nu})$, which gives the inverse metric tensor g^{ab} satisfying

$$g_{ac}g^{cb} = \delta_a{}^b, \tag{48}$$

where $\delta_a{}^b$ is the identity tensor, or equivalently the Kronecker delta in local coordinates.

For any smooth scalar field $f \in C^\infty(\mathcal{M}_L)$, its covariant derivative $\nabla_b f$ is a covector field, i.e., a smooth section of the cotangent bundle $T^*\mathcal{M}_L$. Since the inverse metric g^{ab} is a tensor field of type $(2, 0)$, contracting one index of g^{ab} with the covariant index of $\nabla_b f$ produces a vector field:

$$v^a := g^{ab}\nabla_b f. \tag{49}$$

This operation is well-defined because the repeated index b is contracted, leaving one free contravariant index a. Thus, v^a is a tangent vector field on $\mathcal{M}_L$.

The uniqueness of this raised vector field follows from the uniqueness of the inverse metric. Suppose there exists another vector field $\tilde{v}^a$ satisfying the same metric-dual relation, namely

$$g_{ab}\tilde{v}^a = \nabla_b f. \tag{50}$$

Applying the inverse metric g^{cb} to both sides gives

$$g^{cb}g_{ab}\tilde{v}^a = g^{cb}\nabla_b f. \tag{51}$$

Using $g^{cb}g_{ab} = \delta_a{}^c$, we obtain

$$\tilde{v}^c = g^{cb}\nabla_b f. \tag{52}$$

Therefore, $\tilde{v}^a = v^a$, proving uniqueness.

Now consider a local activation coordinate system $(x^1, \ldots, x^\mu, \ldots, x^d)$ around p. The coordinate basis of the tangent space is $\{(\partial/\partial x^\mu)^a\}_{\mu=1}^d$, and the corresponding dual basis of the cotangent space is $\{(dx^\mu)_a\}_{\mu=1}^d$, satisfying

$$(dx^\mu)_a \left(\frac{\partial}{\partial x^\nu}\right)^a = \delta^\mu_\nu. \tag{53}$$

Hence, the vector field v^a admits the local expansion

$$v^a = v^\mu \left(\frac{\partial}{\partial x^\mu}\right)^a. \tag{54}$$

Applying the dual basis $(dx^\mu)_a$ to both sides yields

$$(dx^\mu)_a v^a = v^\nu (dx^\mu)_a \left(\frac{\partial}{\partial x^\nu}\right)^a = v^\nu \delta^\mu_\nu = v^\mu. \tag{55}$$

Therefore,

$$v^\mu = (dx^\mu)_a v^a. \tag{56}$$

Substituting $v^a = g^{ab}\nabla_b f$ into the above expression gives

$$v^\mu = (dx^\mu)_a g^{ab}\nabla_b f. \tag{57}$$

In the coordinate basis, the components of the inverse metric are $g^{\mu\nu} = (dx^\mu)_a (dx^\nu)_b g^{ab}$, and the components of the covariant derivative are $\nabla_\nu f = (\partial/\partial x^\nu)^b \nabla_b f$. Therefore,

$$v^\mu = g^{\mu\nu}\nabla_\nu f. \tag{58}$$

Since f is a scalar field, its covariant derivative coincides with its ordinary partial derivative in local coordinates, namely

$$\nabla_\nu f = \frac{\partial f}{\partial x^\nu}. \tag{59}$$

Consequently, the raised vector field can be written as

$$v^a = v^\mu \left(\frac{\partial}{\partial x^\mu}\right)^a, \qquad v^\mu = (dx^\mu)_a v^a = g^{\mu\nu}\nabla_\nu f = g^{\mu\nu}\frac{\partial f}{\partial x^\nu}. \tag{60}$$

This completes the proof. □

B.5 Proof of Lemma 3

Lemma 3 (Inter-layer Metric Pullback). *Let $\mathcal{M}_{L-1}$ and $\mathcal{M}_L$ be the representation manifolds of the $(L-1)$-th and L-th layers, respectively. Let $\phi_{L-1} : \mathcal{M}_{L-1} \to \mathcal{M}_L$ be the smooth cross-layer mapping defined in Definition 2. Starting with $(g_{cd})^L$ on $\mathcal{M}_L$, its backward recursion to the previous activation space $\mathcal{M}_{L-1}$ yields the $(0,2)$-tensor:*

$$(g_{ab})^{L-1}\big|_p(v^a, u^b) = (g_{cd})^L\big|_{\phi_{L-1}(p)}\big((\phi_{L-1*}v)^c, (\phi_{L-1*}u)^d\big), \qquad \forall v, u \in T_p\mathcal{M}_{L-1}, \tag{61}$$

where $\phi_{L-1*} : T_p\mathcal{M}_{L-1} \to T_{\phi_{L-1}(p)}\mathcal{M}_L$ denotes the pushforward induced by ϕ_{L-1}.

Proof. The above construction is the pullback of the metric tensor $(g_{cd})^L$ along the smooth mapping ϕ_{L-1}. Define

$$(g^{L-1})_p = \phi_{L-1}^*(g^L)_p.$$

For any $v, u \in T_p\mathcal{M}_{L-1}$, this means

$$(g^{L-1})_p(v,u) = (g^L)_{\phi_{L-1}(p)}\big(\phi_{L-1*}v, \phi_{L-1*}u\big).$$

We first show that this defines a $(0,2)$-tensor at p. Since ϕ_{L-1*} is linear and $(g^L)_{\phi_{L-1}(p)}$ is bilinear, for any scalars α, β and tangent vectors $v_1, v_2, u \in T_p\mathcal{M}_{L-1}$, we have

$$(g^{L-1})_p(\alpha v_1 + \beta v_2, u) = \alpha(g^{L-1})_p(v_1, u) + \beta(g^{L-1})_p(v_2, u).$$

The same argument applies to the second argument. Therefore, $(g^{L-1})_p$ is bilinear and hence is a covariant two-tensor.

Next, we verify its local coordinate expression. Choose local coordinates $(x^1, \ldots, x^m)$ on $\mathcal{M}_{L-1}$ and $(y^1, \ldots, y^n)$ on $\mathcal{M}_L$. Write the coordinate expression of ϕ_{L-1} as

$$y^c = \phi_{L-1}^c(x).$$

Then the pushforward acts on the coordinate basis by

$$\phi_{L-1*}\left(\frac{\partial}{\partial x^a}\right) = \frac{\partial \phi_{L-1}^c}{\partial x^a}\frac{\partial}{\partial y^c}.$$

Therefore, the components of the pulled-back tensor are

$$(g^{L-1})_{ab}(x) = (g^L)_{cd}(\phi_{L-1}(x))\frac{\partial \phi_{L-1}^c}{\partial x^a}\frac{\partial \phi_{L-1}^d}{\partial x^b}. \tag{62}$$

Because ϕ_{L-1} is smooth and $(g^L)_{cd}$ is smooth, the components $(g^{L-1})_{ab}$ are also smooth. Thus g^{L-1} is a smooth $(0,2)$-tensor field on $\mathcal{M}_{L-1}$.

Moreover, since g^L is symmetric, we have

$$(g^{L-1})_p(v,u) = (g^L)_{\phi_{L-1}(p)}(\phi_{L-1*}v, \phi_{L-1*}u) = (g^L)_{\phi_{L-1}(p)}(\phi_{L-1*}u, \phi_{L-1*}v) = (g^{L-1})_p(u,v).$$

Hence the pulled-back tensor is symmetric.

If ϕ_{L-1*} is injective at p and g^L is positive definite, then for any nonzero $v \in T_p\mathcal{M}_{L-1}$, one has $\phi_{L-1*}v \neq 0$, and therefore

$$(g^{L-1})_p(v,v) = (g^L)_{\phi_{L-1}(p)}(\phi_{L-1*}v, \phi_{L-1*}v) > 0.$$

Thus, under this standard local immersion condition, the pulled-back tensor is non-degenerate and defines a Riemannian metric on $\mathcal{M}_{L-1}$. Without this injectivity condition, the construction still defines a smooth symmetric $(0,2)$-tensor, but it may be degenerate. □

C Kantorovich Dual Construction and Properties of the Scalar Field

C.1 Kantorovich-Rubinstein Duality and Saturation Condition

The geometric foundation of our proposition is established within the Riemannian manifold $(\mathcal{M}_L, g_{ab})$ of the L-th layer's activation space. We define two distinct probability measures: μ_L^- representing the source distribution of hallucination states, and μ_L^+ representing the target distribution of truthful states.

To quantify the effort required to rectify hallucinations, we employ the Wasserstein-1 (W_1) distance. In its primal Monge-Kantorovich formulation, it seeks the optimal transport plan $\pi(y,x)$ that minimizes the expected transport cost:

$$W_1(\mu_L^-, \mu_L^+) = \inf_{\pi \in \Pi(\mu_L^-, \mu_L^+)} \int_{\mathcal{M}_L \times \mathcal{M}_L} d_g(y,x)\, d\pi(y,x), \tag{63}$$

where $\Pi(\mu_L^-, \mu_L^+)$ is the set of all joint probability measures whose marginals are μ_L^- and μ_L^+, respectively. The cost function $d_g(y,x)$ is the intrinsic geodesic distance on the manifold. Solving this primal problem requires tracking the combinatorial mass flow between all pairs of states, which is computationally intractable in high-dimensional latent spaces.

To decouple the marginal constraints, we introduce two continuous scalar functions, $f_L(x)$ and $h_L(y)$, acting as Lagrange multipliers for the target measure μ_L^+ and source measure μ_L^-, respectively. The constrained optimization is thus rewritten via the Lagrangian $\mathcal{L}$:

$$\inf_{\pi \geq 0} \sup_{f_L, h_L} \Bigg[\int d_g(y,x)\, d\pi(y,x) + \int f_L(x)(d\mu_L^+(x) - d\pi(y,x)) + \int h_L(y)(d\mu_L^-(y) - d\pi(y,x)) \Bigg]. \tag{64}$$

By applying the Minimax theorem to swap the infimum and supremum, and subsequently grouping the terms integrated with respect to $d\pi(y,x)$, the objective becomes:

$$\sup_{f_L, h_L} \Bigg[\int_{\mathcal{M}_L} f_L(x) d\mu_L^+(x) + \int_{\mathcal{M}_L} h_L(y) d\mu_L^-(y) + \inf_{\pi \geq 0} \int_{\mathcal{M}_L \times \mathcal{M}_L} (d_g(y,x) - f_L(x) - h_L(y))\, d\pi(y,x) \Bigg]. \tag{65}$$

For the inner infimum to not diverge to $-\infty$, the integrand must be non-negative everywhere. This strictly requires that $f_L(x) + h_L(y) \leq d_g(y,x)$ for all $x, y \in \mathcal{M}_L$.

Because the cost function $d_g(y,x)$ is a valid Riemannian metric (satisfying non-negativity, symmetry, and the triangle inequality), we can simplify the relationship between the dual variables f_L and h_L. To maximize the supremum, we want $h_L(y)$ to be as large as possible subject

to the constraint $h_L(y) \leq d_g(y,x) - f_L(x)$. This optimal $h_L(y)$ is given by the c-transform: $f_L^c(y) = \inf_x\{d_g(y,x) - f_L(x)\}$.

Since $d_g(y,y) = 0$, evaluating the infimum at $x = y$ yields $h_L(y) \leq -f_L(y)$. Due to the triangle inequality $d_g(y,x) \leq d_g(y,z) + d_g(z,x)$, the optimal choice elegantly collapses to $h_L(y) = -f_L(y)$. Substituting this back into our constraint yields the fundamental condition:

$$f_L(x) - f_L(y) \leq d_g(y,x). \tag{66}$$

This inequality reveals that the optimal dual scalar field f_L must be a 1-Lipschitz function over the manifold. Consequently, the dual problem simplifies to its final Kantorovich-Rubinstein form:

$$W_1(\mu_L^-, \mu_L^+) = \sup_{f_L \in \mathrm{Lip}_1(\mathcal{M}_L)} \left(\int_{\mathcal{M}_L} f_L d\mu_L^+ - \int_{\mathcal{M}_L} f_L d\mu_L^- \right). \tag{67}$$

The feasible region of this dual problem, $\mathrm{Lip}_1(\mathcal{M}_L)$, enforces a strict geometric limit on the behavior of f_L: The macro-scale constraint $|f_L(x) - f_L(y)| \leq d_g(y,x)$ guarantees that the difference in the scalar field between any two states cannot exceed their actual shortest-path distance. By examining this global bound within an infinitesimal neighborhood along a unit-speed geodesic $\gamma(t)$, we take the limit as the distance $\Delta t \to 0$:

$$\lim_{\Delta t \to 0} \frac{|f_L(\gamma(\Delta t)) - f_L(\gamma(0))|}{\Delta t} \leq 1. \tag{68}$$

Applying the chain rule on the manifold, this limit represents the directional derivative, ensuring $|\langle \nabla_a f_L, \dot{\gamma}^a \rangle_g| \leq 1$. Since this holds for all directions $\dot{\gamma}^a$, the supremum of this directional derivativewhich is the Riemannian norm of the gradientis globally bounded:

$$\|\nabla_a f_L\|_g = \sqrt{g^{ab}(\nabla_a f_L)(\nabla_b f_L)} \leq 1. \tag{69}$$

Now, let f_L^* be the scalar field that achieves this supremum:

$$f_L^* = \underset{f_L \in \mathrm{Lip}_1(\mathcal{M}_L)}{\arg\max} \left[\int_{\mathcal{M}_L} f_L(x) d\mu_L^+(x) - \int_{\mathcal{M}_L} f_L(x) d\mu_L^-(x) \right]. \tag{70}$$

To rigorously establish the saturation of the geometric bound, we invoke the strong duality of the Wasserstein-1 metric. At optimality, the minimum of the primal transport cost exactly equals the maximum of the dual expectation gap. Let π^* be the optimal transport plan and f_L^* be the optimal dual scalar field; strong duality dictates:

$$\int_{\mathcal{M}_L \times \mathcal{M}_L} d_g(y,x)\, d\pi^*(y,x) = \int_{\mathcal{M}_L} f_L^*(x) d\mu_L^+(x) - \int_{\mathcal{M}_L} f_L^*(y) d\mu_L^-(y). \tag{71}$$

Because the marginals of the optimal transport plan $\pi^*(y,x)$ are exactly defined as $\mu_L^-(y)$ and $\mu_L^+(x)$, we can perfectly express the right-hand side as a joint integral with respect to $d\pi^*(y,x)$:

$$\int_{\mathcal{M}_L \times \mathcal{M}_L} d_g(y,x)\, d\pi^*(y,x) = \int_{\mathcal{M}_L \times \mathcal{M}_L} (f_L^*(x) - f_L^*(y))\, d\pi^*(y,x). \tag{72}$$

Rearranging the terms yields a single unified integral:

$$\int_{\mathcal{M}_L \times \mathcal{M}_L} \Big[d_g(y,x) - (f_L^*(x) - f_L^*(y)) \Big]\, d\pi^*(y,x) = 0. \tag{73}$$

Recall that the global 1-Lipschitz constraint necessitates $f_L^*(x) - f_L^*(y) \leq d_g(y,x)$ for all $x, y \in \mathcal{M}_L$. Therefore, the bracketed integrand is strictly non-negative everywhere. In measure theory, if the integral of a non-negative function evaluates to zero, the function itself must be zero almost everywhere on the support of the measure. Thus, for any pair of states (y,x) in the support of the optimal transport plan $\mathrm{spt}(\pi^*)$, the Duality Saturation Condition must hold strictly:

$$f_L^*(x) - f_L^*(y) = d_g(y,x). \tag{74}$$

C.2 Proof of Theorem: Equivalence Between Geodesics and Scalar Field Tangent Vectors

Theorem 1 (Equivalence Between Geodesics and Scalar Field Tangent Vectors). *Let $\gamma : [0, \ell] \to \mathcal{M}_L$ be a geodesic connecting a hallucinated activation y and a truthful activation x, with $(y, x) \in \mathrm{supp}(\pi^*)$, where π^* denotes the optimal transport plan. According to the duality saturation in Eq. 74, together with the line integral theorem and geodesic length definition, we obtain:*

$$f_L^*(x) - f_L^*(y) = \int_0^\ell \left\langle g^{ab} \nabla_b f_L^*(\gamma(t)), \dot{\gamma}(t) \right\rangle_g \, dt = \int_0^\ell \|\dot{\gamma}(t)\|_g \, dt = d_g(x, y). \tag{75}$$

where $\dot{\gamma}(t)$ represents the tangent vector of the geodesic at the point $\gamma(t)$ with respect to the geodesic parameter t, i.e., $\dot{\gamma}(t) = \frac{d\gamma(t)}{dt}$. On the other hand, by the gradient-norm bound in Eq. 3 and the Cauchy–Schwarz inequality, the directional derivative along γ satisfies the pointwise upper bound

$$\left\langle g^{ab} \nabla_b f_L^*(\gamma(t)), \dot{\gamma}(t) \right\rangle_g \le \left\| g^{ab} \nabla_b f_L^*(\gamma(t)) \right\|_g \|\dot{\gamma}(t)\|_g \le \|\dot{\gamma}(t)\|_g. \tag{76}$$

Combining Eq. 75 and Eq. 76, the upper bound in Eq. 76 must be attained almost everywhere along γ; otherwise, the integral on the left-hand side of Eq. 75 would be strictly smaller than the total geodesic length. Therefore, the equality conditions of the Cauchy–Schwarz inequality and the gradient-norm bound hold almost everywhere, yielding

$$g^{ab} \nabla_b f_L^*(\gamma(t)) = \frac{\dot{\gamma}^a(t)}{\|\dot{\gamma}(t)\|_g} \qquad \text{for a.e. } t \in [0, \ell]. \tag{77}$$

According to Lemma 2, $g^{ab} \nabla_b f_L^*(\gamma(t))$ constitutes the tangent vector v^a of f_L^*. The detailed proof and derivation are provided in Appendix C.2.

Proof. Let $\gamma : [0, \ell] \to \mathcal{M}_L$ be a geodesic connecting the hallucinated state y to the truthful state x, satisfying $\gamma(0) = y$ and $\gamma(\ell) = x$. According to the definition of curve length in Definition 1, the length of γ is given by

$$L_g(\gamma) = \int_0^\ell \|\dot{\gamma}(t)\|_g \, dt. \tag{78}$$

Since γ is chosen as a geodesic realizing the intrinsic distance between y and x, its length coincides with the geodesic distance:

$$L_g(\gamma) = d_g(y, x) = \int_0^\ell \|\dot{\gamma}(t)\|_g \, dt. \tag{79}$$

This identity expresses the total mitigation cost from the hallucinated activation state y to the truthful activation state x as the accumulated Riemannian speed along the geodesic path.

Next, we examine how the scalar field f_L^* changes along the same geodesic. By applying the fundamental theorem of calculus along the smooth curve $\gamma(t)$, the variation of the scalar field between the two endpoints can be written as

$$f_L^*(x) - f_L^*(y) = f_L^*(\gamma(\ell)) - f_L^*(\gamma(0)) = \int_0^\ell \frac{d}{dt} f_L^*(\gamma(t)) \, dt. \tag{80}$$

Using the chain rule on the Riemannian manifold, the derivative of the scalar field along the curve is

$$\frac{d}{dt} f_L^*(\gamma(t)) = \nabla_a f_L^*(\gamma(t)) \, \dot{\gamma}^a(t). \tag{81}$$

Here, $\nabla_a f_L^*$ is a covector field, while $\dot{\gamma}^a(t)$ is the tangent vector of the geodesic. According to the index-raising operation in Lemma 2, the covector field $\nabla_b f_L^*$ can be uniquely raised to the tangent vector field

$$v^a(\gamma(t)) = g^{ab} \nabla_b f_L^*(\gamma(t)). \tag{82}$$

Together with the Riemannian inner product defined in Definition 1, the contraction in Eq. 81 can equivalently be written as

$$\nabla_a f_L^*(\gamma(t)) \, \dot{\gamma}^a(t) = \left\langle g^{ab} \nabla_b f_L^*(\gamma(t)), \dot{\gamma}(t) \right\rangle_g. \tag{83}$$

Indeed, by the definition of the Riemannian inner product, we have

$$\begin{aligned}\left\langle g^{ab}\nabla_b f_L^*(\gamma(t)), \dot\gamma(t)\right\rangle_g &= g_{ac}\left(g^{ab}\nabla_b f_L^*(\gamma(t))\right)\dot\gamma^c(t)\\ &= \delta_c{}^b\nabla_b f_L^*(\gamma(t))\dot\gamma^c(t)\\ &= \nabla_c f_L^*(\gamma(t))\dot\gamma^c(t),\end{aligned} \tag{84}$$

where $g_{ac}g^{ab} = \delta_c{}^b$ follows from the non-degeneracy of the metric tensor.

Substituting Eq. 83 into Eq. 80, we obtain the line-integral expression of the scalar field variation:

$$f_L^*(x) - f_L^*(y) = \int_0^\ell \left\langle g^{ab}\nabla_b f_L^*(\gamma(t)), \dot\gamma(t)\right\rangle_g dt. \tag{85}$$

We now connect this line-integral expression with the Kantorovich–Rubinstein duality saturation condition. According to the duality saturation condition in Eq. 74, for any transported pair $(y, x) \in \mathrm{spt}(\pi^*)$, the scalar field satisfies

$$f_L^*(x) - f_L^*(y) = d_g(y, x). \tag{86}$$

Combining Eq. 85 with Eq. 86, we have

$$\int_0^\ell \left\langle g^{ab}\nabla_b f_L^*(\gamma(t)), \dot\gamma(t)\right\rangle_g dt = d_g(y, x). \tag{87}$$

Further combining Eq. 87 with the geodesic length identity in Eq. 79, we obtain

$$\int_0^\ell \left\langle g^{ab}\nabla_b f_L^*(\gamma(t)), \dot\gamma(t)\right\rangle_g dt = \int_0^\ell \|\dot\gamma(t)\|_g\, dt. \tag{88}$$

This equation states that, along the geodesic, the accumulated directional increase of the scalar field exactly equals the accumulated geodesic length.

On the other hand, since f_L^* is a feasible solution of the dual problem, it satisfies the 1-Lipschitz constraint on $\mathcal{M}_L$. According to the equivalent local gradient-norm characterization in Eq. 69, for differentiable points on the manifold, we have

$$\left\|g^{ab}\nabla_b f_L^*(\gamma(t))\right\|_g \le 1 \qquad \text{for a.e. } t \in [0, \ell]. \tag{89}$$

By the Cauchy–Schwarz inequality under the Riemannian metric, we further have

$$\begin{aligned}\left\langle g^{ab}\nabla_b f_L^*(\gamma(t)), \dot\gamma(t)\right\rangle_g &\le \left\|g^{ab}\nabla_b f_L^*(\gamma(t))\right\|_g \|\dot\gamma(t)\|_g\\ &\le \|\dot\gamma(t)\|_g,\end{aligned} \tag{90}$$

for almost every $t \in [0, \ell]$.

Eq. 88 shows that the integral of the directional derivative induced by the scalar field exactly equals the total geodesic length. However, Eq. 90 shows that the integrand is pointwise upper-bounded by $\|\dot\gamma(t)\|_g$. Therefore, this upper bound must be attained almost everywhere along γ. Otherwise, if the inequality were strict on a subset of $[0, \ell]$ with positive measure, then the integral of the left-hand side in Eq. 88 would be strictly smaller than the total geodesic length, which contradicts Eq. 88. Consequently, we obtain

$$\left\langle g^{ab}\nabla_b f_L^*(\gamma(t)), \dot\gamma(t)\right\rangle_g = \|\dot\gamma(t)\|_g \qquad \text{for a.e. } t \in [0, \ell]. \tag{91}$$

The equality in Eq. 91 implies that both inequalities in Eq. 90 are saturated almost everywhere. First, the 1-Lipschitz gradient bound must be tight:

$$\left\|g^{ab}\nabla_b f_L^*(\gamma(t))\right\|_g = 1 \qquad \text{for a.e. } t \in [0, \ell]. \tag{92}$$

Second, the equality condition of the Cauchy–Schwarz inequality requires that the tangent vector of f_L^* and the geodesic tangent vector $\dot\gamma(t)$ be positively aligned. Therefore, the tangent vector must coincide with the normalized geodesic tangent direction:

$$g^{ab}\nabla_b f_L^*(\gamma(t)) = \frac{\dot\gamma^a(t)}{\|\dot\gamma(t)\|_g} \qquad \text{for a.e. } t \in [0, \ell]. \tag{93}$$

Therefore, along the geodesic from hallucinated activation states to truthful activation states, the tangent vector of the scalar field f_L^* is equivalent to the normalized tangent direction of the mitigation path. This establishes the geometric equivalence between the Kantorovich-dual scalar field and the local mitigation direction on the activation manifold.

□

C.3 Proof of Corollary: Scalar Field Based Identification of H-Neurons

Corollary 1 (Scalar Field Based Identification of H-Neurons). *For a local activation state $\gamma(t)$ along the geodesic, since f_L^* is a smooth scalar field defined on the activation manifold, its covariant derivative can be expanded as $\nabla_a f_L^* = \sum_{\mu=1}^{d} (\nabla_\mu f_L^*)(dx^\mu)_a, \nabla_\mu f_L^* = \left(\frac{\partial}{\partial x^\mu}\right)^a \nabla_a f_L^* = \frac{\partial f_L^*}{\partial x^\mu}$. Applying the equality in Theorem 1 to the μ-th neuron activation, and combining it with the tangent vector computation in Lemma 2 and the preceding expansion, we obtain*

$$v^\mu(t) = g^{\mu\nu}\nabla_\nu f_L^*(\gamma(t)) = g^{\mu\nu}\frac{\partial f_L^*}{\partial x^\nu}(\gamma(t)) = \frac{1}{\|\dot\gamma(t)\|_g}\frac{d\gamma^\mu(t)}{dt}. \tag{94}$$

*Here, $v^\mu(t)$ denotes the μ-th component of the tangent direction induced by f_L^*. Since $\frac{d\gamma^\mu(t)}{dt} = \|\dot\gamma(t)\|_g v^\mu(t)$, and $\|\dot\gamma(t)\|_g > 0$ for a regular geodesic, **the sign of $v^\mu(t)$ directly determines whether the μ-th activation is locally enhanced or suppressed along the mitigation direction**. Detailed explanations and proofs can be found in the Appendix C.3.*

- ***Hallucination-Associated Neurons (H-Neurons)*** *($v^\mu = g^{\mu\nu}\nabla_\nu f_L^* = g^{\mu\nu}\frac{\partial f_L^*}{\partial x^\nu} < 0$): A negative local tangent component indicates that the μ-th activation is locally decreasing along the geodesic. This neuron corresponds to a hallucination-direction whose activation is locally suppressed during the transition toward the truthful state. A more comprehensive description of H-Neurons can be found in Definition 3.*

Proof. We refer to Eq. 76 as the *equivalence of tangent vectors*. It shows that, along the geodesic, the metric-dual vector field induced by the covariant derivative $\nabla_a f_L^*$ coincides with the normalized tangent direction of the mitigation trajectory:

$$g^{ab}\nabla_b f_L^*(\gamma(t)) = \frac{\dot\gamma^a(t)}{\|\dot\gamma(t)\|_g} \qquad \text{for a.e. } t \in [a,b]. \tag{95}$$

Different from the unit-speed case, the right-hand side is not the raw tangent vector $\dot\gamma^a(t)$, but its normalized direction. Thus, the scalar field f_L^* determines the geometric direction of the shortest mitigation path, while the actual activation variation additionally depends on the geodesic speed $\|\dot\gamma(t)\|_g$.

To express this gradient-guided mitigation dynamics at the neuron level, we represent the activation state on $\mathcal{M}_L$ as

$$\gamma(t) = (\gamma^1(t), \ldots, \gamma^d(t)), \tag{96}$$

where $\gamma^\mu(t)$ denotes the activation value of the μ-th neuron at the local state $\gamma(t)$.

To strictly decompose the mitigation direction, we consider the general formula for the covariant derivative of a (k,l)-tensor in Lemma 1:

$$\nabla_a T^{b_1\ldots b_k}{}_{c_1\ldots c_l} = \partial_a T^{b_1\ldots b_k}{}_{c_1\ldots c_l} + \sum_{i=1}^{k}\Gamma^{b_i}_{ao}T^{b_1\ldots o\ldots b_k}{}_{c_1\ldots c_l} - \sum_{j=1}^{l}\Gamma^{o}_{ac_j}T^{b_1\ldots b_k}{}_{c_1\ldots o\ldots c_l}. \tag{97}$$

Since f_L^* is a smooth scalar field defined on the activation manifold (a $(0,0)$-tensor), $k=0$ and $l=0$. Thus, all terms involving the Christoffel symbols Γ vanish, and the covariant derivative naturally reduces to the standard partial derivative, i.e., $\nabla_a f_L^* = \partial_a f_L^*$.

Refer to Section B.2.2, as a 1-form, $\nabla_a f_L^*$ can be expanded in the local coordinate basis $\{dx^\nu\}$ of the cotangent space:

$$\nabla_a f_L^* = \sum_{\nu=1}^{d}(\nabla_\nu f_L^*)(dx^\nu)_a. \tag{98}$$

Refer to Eq. 33, applying the dual basis vector $\left(\frac{\partial}{\partial x^\mu}\right)^a$ to both sides yields the inner product. By the definition of the dual basis, the contraction of a basis vector and a dual basis vector yields the Kronecker delta: $(dx^\nu)_a \left(\frac{\partial}{\partial x^\mu}\right)^a = \delta^\nu_\mu$. This selectively collapses the summation:

$$\left(\frac{\partial}{\partial x^\mu}\right)^a \nabla_a f_L^* = \sum_{\nu=1}^{d} (\nabla_\nu f_L^*) \left[(dx^\nu)_a \left(\frac{\partial}{\partial x^\mu}\right)^a\right] = \sum_{\nu=1}^{d} (\nabla_\nu f_L^*)\delta^\nu_\mu = \nabla_\mu f_L^*. \tag{99}$$

Therefore, we rigorously obtain $\nabla_\mu f_L^* = \frac{\partial f_L^*}{\partial x^\mu}$.

Applying the equivalence of tangent vectors to the μ-th activation dimension gives

$$g^{\mu\nu}\nabla_\nu f_L^*(\gamma(t)) = \frac{\dot\gamma^\mu(t)}{\|\dot\gamma(t)\|_g} = \frac{1}{\|\dot\gamma(t)\|_g}\frac{d\gamma^\mu(t)}{dt}. \tag{100}$$

Accordingly, in light of Lemma 2, we define the local tangent vector component induced by the scalar field as

$$v^\mu(t) := g^{\mu\nu}\nabla_\nu f_L^*(\gamma(t)) = g^{\mu\nu}\frac{\partial f_L^*}{\partial x^\nu}(\gamma(t)). \tag{101}$$

Here, $v^\mu(t)$ denotes the μ-th component of the normalized tangent direction induced by f_L^* at the current activation state. Equivalently, the instantaneous activation dynamics satisfy

$$\frac{d\gamma^\mu(t)}{dt} = \|\dot\gamma(t)\|_g v^\mu(t). \tag{102}$$

To explicitly observe the local variation of the activation value, we consider an infinitesimal transition step $\Delta t > 0$ along the geodesic. Using a first-order Taylor expansion, the change in the μ-th neuron's activation is:

$$\Delta\gamma^\mu(t) \approx \gamma^\mu(t+\Delta t) - \gamma^\mu(t) \approx \frac{d\gamma^\mu(t)}{dt}\Delta t = \|\dot\gamma(t)\|_g v^\mu(t)\Delta t. \tag{103}$$

Since γ is a regular geodesic, the geodesic speed is strictly positive ($\|\dot\gamma(t)\|_g > 0$). Given $\Delta t > 0$, the scalar factor $\|\dot\gamma(t)\|_g\Delta t$ is always strictly positive. Consequently, the sign of the local activation variation $\Delta\gamma^\mu(t)$ is entirely determined by the sign of $v^\mu(t)$. This geometric criterion directly indicates whether the μ-th activation is locally enhanced or suppressed along the mitigation direction.

For notational simplicity, we omit the explicit dependence on $\gamma(t)$ in the following neuron-wise interpretation. All quantities are evaluated at the current activation state.

- **Hallucination-Associated Neurons (H-Neurons)** ($v^\mu = g^{\mu\nu}\nabla_\nu f_L^* = g^{\mu\nu}\frac{\partial f_L^*}{\partial x^\nu} < 0$): A negative local tangent component indicates that $\Delta\gamma^\mu(t) < 0$. The μ-th activation is locally decreasing along the geodesic. This neuron corresponds to a hallucination-associated direction whose activation is locally suppressed during the transition toward truthfulness.

Thus, the sign of the tangent vector component v^μ provides a local geometric criterion for distinguishing truthfulness-associated and hallucination-associated neuron directions. □

C.4 Proof of Corollary: Monotonicity of the Scalar Field along Geodesics

Corollary 2 (Monotonicity of the Scalar Field along Geodesics). *Let y denote a hallucinated state and x denote a truthful state. Consider the geodesic $\gamma : [0, \ell] \to \mathcal{M}_L$ satisfying $\gamma(0) = y$ and $\gamma(\ell) = x$. The monotonic increase can be further verified by differentiating f_L^* along the geodesic $\gamma(t)$. Since the metric g_{ab} is positive-definite, its inverse g^{ab} is also positive-definite, and thus*

$$\frac{d}{dt}f_L^*(\gamma(t)) = \nabla_a f_L^*\frac{d\gamma^a}{dt} = \|\dot\gamma(t)\|_g \nabla_a f_L^* \left(g^{ab}\nabla_b f_L^*\right) = \|\dot\gamma(t)\|_g \|\nabla f_L^*\|_g^2 \geq 0. \tag{104}$$

The nonnegative derivative shows that f_L^* is non-decreasing along the transport flow from the hallucinated state y to the truthful state x. ***Consequently, for any two states $x_1, x_2 \in \mathcal{M}_L$, if $f_L^*(x_1) > f_L^*(x_2)$, then x_1 is therefore more geometrically aligned with the truthful distribution.*** *Detailed explanations and proofs can be found in the Appendix C.4.*

Proof. Let $\gamma : [0, \ell] \to \mathcal{M}_L$ be the geodesic connecting the hallucinated state $y = \gamma(0)$ to the truthful state $x = \gamma(\ell)$. Unlike the unit-speed case, γ is not necessarily parameterized by arc length. Therefore, its length is given by

$$L_g(\gamma) = \int_0^\ell \|\dot{\gamma}(t)\|_g \, dt. \tag{105}$$

By the fundamental theorem of calculus along the curve, the total variation of the scalar field is

$$f_L^*(x) - f_L^*(y) = \int_0^\ell \frac{d}{dt} f_L^*(\gamma(t)) \, dt. \tag{106}$$

The instantaneous rate of change of f_L^* along γ is given by the chain rule:

$$\frac{d}{dt} f_L^*(\gamma(t)) = \nabla_a f_L^*(\gamma(t)) \dot{\gamma}^a(t). \tag{107}$$

According to the equivalence of tangent vectors in Eq. 77, we have

$$g^{ab} \nabla_b f_L^*(\gamma(t)) = \frac{\dot{\gamma}^a(t)}{\|\dot{\gamma}(t)\|_g} \qquad \text{for a.e. } t \in [0, \ell]. \tag{108}$$

Equivalently,

$$\dot{\gamma}^a(t) = \|\dot{\gamma}(t)\|_g g^{ab} \nabla_b f_L^*(\gamma(t)). \tag{109}$$

We now express the above variation at the neuron-group level. In activation coordinates, the time derivative can be decomposed as

$$\frac{d}{dt} f_L^*(\gamma(t)) = \sum_{\mu=1}^{d} \nabla_\mu f_L^*(\gamma(t)) \frac{d\gamma^\mu(t)}{dt}. \tag{110}$$

From the equivalence of tangent vectors, the coordinate-wise velocity satisfies

$$\frac{d\gamma^\mu(t)}{dt} = \|\dot{\gamma}(t)\|_g g^{\mu\nu} \nabla_\nu f_L^*(\gamma(t)). \tag{111}$$

Substituting this expression gives

$$\frac{d}{dt} f_L^*(\gamma(t)) = \|\dot{\gamma}(t)\|_g \sum_{\mu=1}^{d} \nabla_\mu f_L^*(\gamma(t)) \left(g^{\mu\nu} \nabla_\nu f_L^*(\gamma(t)) \right). \tag{112}$$

The summation is precisely the squared Riemannian norm of the covariant gradient:

$$\sum_{\mu=1}^{d} \nabla_\mu f_L^* \left(g^{\mu\nu} \nabla_\nu f_L^* \right) = g^{\mu\nu} \nabla_\mu f_L^* \nabla_\nu f_L^* = \|\nabla f_L^*\|_g^2. \tag{113}$$

Because the inverse metric $g^{\mu\nu}$ is positive-definite, this quadratic form is nonnegative. Moreover, the equivalence of tangent vectors implies the saturation of the gradient-norm bound along the geodesic, namely

$$\|\nabla f_L^*(\gamma(t))\|_g = 1 \qquad \text{for a.e. } t \in [0, \ell]. \tag{114}$$

Let $\mathcal{N}^+$ and $\mathcal{N}^-$ denote the index sets of T-Neurons and H-Neurons, respectively. Ignoring neutral components, the scalar field variation can be additively decomposed into the contributions of these two neuron groups:

$$\begin{aligned} f_L^*(x) - f_L^*(y) &= \int_0^\ell \|\dot{\gamma}(t)\|_g \left[\sum_{i \in \mathcal{N}^+} \nabla_i f_L^* \left(g^{i\nu} \nabla_\nu f_L^* \right) + \sum_{j \in \mathcal{N}^-} \nabla_j f_L^* \left(g^{j\nu} \nabla_\nu f_L^* \right) \right] dt \\ &= \int_0^\ell \|\dot{\gamma}(t)\|_g \|\nabla f_L^*(\gamma(t))\|_g^2 \, dt \\ &= \int_0^\ell \|\dot{\gamma}(t)\|_g \, dt \\ &= d_g(y, x). \end{aligned} \tag{115}$$

This identity shows that the microscopic neuron-wise contributions collectively explain the macroscopic increase of the scalar field from y to x.

Equivalently, at the infinitesimal level, the monotonic variation of the scalar field along the optimal transport geodesic can be written as

$$\begin{aligned}\frac{d}{dt}f_L^*(\gamma(t)) &= \nabla_a f_L^*(\gamma(t))\frac{d\gamma^a(t)}{dt} \\ &= \|\dot\gamma(t)\|_g\, \nabla_a f_L^*(\gamma(t))\left(g^{ab}\nabla_b f_L^*(\gamma(t))\right) \\ &= \|\dot\gamma(t)\|_g\, \|\nabla f_L^*(\gamma(t))\|_g^2 \geq 0.\end{aligned} \tag{116}$$

Thus, f_L^* is non-decreasing along the transport flow from the hallucinated state y to the truthful state x. Moreover,

$$f_L^*(x) - f_L^*(y) = \int_0^\ell \|\dot\gamma(t)\|_g\, dt = L_g(\gamma) = d_g(y,x), \tag{117}$$

which is exactly consistent with the duality saturation condition in Eq. 5.

Therefore, since

$$f_L^*(x) - f_L^*(y) = d_g(y,x) > 0, \tag{118}$$

the truthful state x receives a larger scalar field value than the hallucinated state y. Equivalently, along the optimal transport flow, the scalar field increases monotonically from hallucinated representations toward truthful representations.

□

C.5 Robustness of the Scalar Field

Proposition 3 (**Robustness of the Scalar Field**). *The scalar field f_L^* is robust both geometrically and statistically. Let $x \in \mathcal{M}_L$ denote an activation state and let $\tilde{x} \in \mathcal{M}_L$ denote its perturbed counterpart. Since f_L^* is 1-Lipschitz with respect to the geodesic distance d_g, we have*

$$|f_L^*(\tilde{x}) - f_L^*(x)| \leq d_g(\tilde{x}, x). \tag{119}$$

Therefore, perturbations of activation states induce only bounded changes in the scalar field value.

Statistically, let μ_L^+ and μ_L^- denote the truthful and hallucinated activation distributions, and let $\hat{\mu}_n^+$ and $\hat{\mu}_m^-$ denote their empirical estimates constructed from n and m samples, respectively. Define

$$\varepsilon_n^+ = \sup_{\mathrm{Lip}_{d_g}(f)\leq 1} \left| \int f\, d\hat{\mu}_n^+ - \int f\, d\mu_L^+ \right|, \tag{120}$$

$$\varepsilon_m^- = \sup_{\mathrm{Lip}_{d_g}(f)\leq 1} \left| \int f\, d\hat{\mu}_m^- - \int f\, d\mu_L^- \right|. \tag{121}$$

Then,

$$\mathbb{E}\left|W_1(\hat{\mu}_n^+, \hat{\mu}_m^-) - W_1(\mu_L^+, \mu_L^-)\right| \leq \mathbb{E}\varepsilon_n^+ + \mathbb{E}\varepsilon_m^-. \tag{122}$$

The two empirical approximation errors are controlled by the statistical complexity of the 1-Lipschitz function class on $\mathcal{M}_L$, which depends on geometric properties of the activation manifold, including its intrinsic complexity. Therefore, when $\mathcal{M}_L$ has controlled intrinsic complexity, the empirical approximation remains stable. Consequently, f_L^* is stable under local perturbations and its empirical estimation remains statistically controlled.

Proof. The robustness of the scalar field f_L^* follows from two complementary properties: geometric stability induced by the Lipschitz constraint and statistical stability of the empirical Wasserstein objective.

Local robustness under semantic perturbations. Since f_L^* is the optimizer of the Kantorovich–Rubinstein dual formulation, it belongs to the class of 1-Lipschitz functions on the Riemannian manifold $(\mathcal{M}_L, g_{bc})$:

$$|f_L^*(x) - f_L^*(p)| \leq d_g(x,p), \qquad \forall x, p \in \mathcal{M}_L. \tag{123}$$

Equivalently, whenever differentiable,

$$g^{ab}\nabla_a f_L^* \nabla_b f_L^* \leq 1, \tag{124}$$

Let $\tilde{x}$ be a perturbed activation state satisfying$d_g(x,\tilde{x}) \leq \|\epsilon\|_g$. By Lipschitz continuity,

$$|f_L^*(\tilde{x}) - f_L^*(x)| \leq d_g(\tilde{x}, x) \leq \|\epsilon\|_g. \tag{125}$$

Hence, small intrinsic perturbations of activation states can only produce proportionally small changes in the scalar field value.

Furthermore, consider a threshold τ and define the margin

$$m(x) = |f_L^*(x) - \tau|. \tag{126}$$

If $d_g(x,\tilde{x}) < m(x)$,then

$$|f_L^*(\tilde{x}) - f_L^*(x)| < |f_L^*(x) - \tau|. \tag{127}$$

Therefore, the sign of $f_L^*(x) - \tau$ cannot change under such perturbations. This proves the local geometric robustness of the scalar field-based indicator.

Statistical robustness under approximation of finite sample. For simplicity, we temporarily omit the layer index L. Assume that the involved distributions have finite first moments and that the Kantorovich–Rubinstein duality applies.

By the triangle inequality of Wasserstein distance,

$$\begin{aligned} W_1(\hat{\mu}_n^+, \hat{\mu}_m^-) \leq & W_1(\hat{\mu}_n^+, \mu^+) + W_1(\mu^+, \mu^-) \\ & + W_1(\mu^-, \hat{\mu}_m^-). \end{aligned} \tag{128}$$

Similarly,

$$\begin{aligned} W_1(\mu^+, \mu^-) \leq & W_1(\mu^+, \hat{\mu}_n^+) + W_1(\hat{\mu}_n^+, \hat{\mu}_m^-) \\ & + W_1(\hat{\mu}_m^-, \mu^-). \end{aligned} \tag{129}$$

Combining these two inequalities gives

$$\left|W_1(\hat{\mu}_n^+, \hat{\mu}_m^-) - W_1(\mu^+, \mu^-)\right| \leq W_1(\hat{\mu}_n^+, \mu^+) + W_1(\hat{\mu}_m^-, \mu^-). \tag{130}$$

Using the Kantorovich–Rubinstein duality,

$$W_1(\hat{\mu}_n^+, \mu^+) = \sup_{\mathrm{Lip}_{d_g}(f) \leq 1} \left| \int f \, d\hat{\mu}_n^+ - \int f \, d\mu^+ \right| = \varepsilon_n^+, \tag{131}$$

and$W_1(\hat{\mu}_m^-, \mu^-) = \varepsilon_m^-$.Hence,

$$\left|W_1(\hat{\mu}_n^+, \hat{\mu}_m^-) - W_1(\mu^+, \mu^-)\right| \leq \varepsilon_n^+ + \varepsilon_m^-. \tag{132}$$

Taking expectations yields

$$\mathbb{E}\left|W_1(\hat{\mu}_n^+, \hat{\mu}_m^-) - W_1(\mu^+, \mu^-)\right| \leq \mathbb{E}\varepsilon_n^+ + \mathbb{E}\varepsilon_m^-. \tag{133}$$

The two terms correspond to uniform empirical approximation errors over the 1-Lipschitz function class. Their magnitudes are controlled by the complexity of $\mathrm{Lip}_{d_g}(\mathcal{M}_L)$, such as metric entropy or Rademacher complexity. Since these quantities depend on the intrinsic geometric structure of $\mathcal{M}_L$, rather than only the ambient dimension, the empirical approximation remains controlled when the activation manifold has low intrinsic complexity.

□

D Proof of Pullback of Scalar Fields Across Layers

D.1 Proof of Proposition: Recursive Construction via Normalized Pullback

Proposition 1 (Recursive Construction via Normalized Pullback). *Consider the cross-layer mapping $\phi_{L-1} : \mathcal{M}_{L-1} \to \mathcal{M}_L$ as defined in Definition 2. Assume that ϕ_{L-1} satisfies the K_{L-1}-Lipschitz continuity in Assumption 2, and let f_L^* be the optimal 1-Lipschitz scalar field on $\mathcal{M}_L$. Then a candidate scalar field on $\mathcal{M}_{L-1}$ can be induced by the normalized pullback operator:*

$$f_{L-1}(H^{(L-1)}) = \frac{1}{K_{L-1}}(\phi_{L-1}^* f_L^*)(H^{(L-1)}) = \frac{1}{K_{L-1}} f_L^*\Big(\phi_{L-1}(H^{(L-1)})\Big), \tag{134}$$

where ϕ_{L-1}^* denotes the pullback operator that transports the scalar field from layer L to layer $L-1$. In particular, the pullback satisfies $\phi_{L-1}^* f = f \circ \phi_{L-1}$.

Proof. On the manifold $\mathcal{M}_L$, according to Kantorovich duality, the Wasserstein-1 distance $W_1(\mu_L^+, \mu_L^-)$ is defined as the supremum of the difference in integrals over all 1-Lipschitz functions. Let $f_L^* \in \mathrm{Lip}_1(\mathcal{M}_L)$ be the scalar field function that achieves this supremum, satisfying

$$|f_L^*(H_a^{(L)}) - f_L^*(H_b^{(L)})| \leq d_{g^{(L)}}(H_a^{(L)}, H_b^{(L)}), \tag{135}$$

for any $H_a^{(L)}, H_b^{(L)} \in \mathcal{M}_L$.

Consider the parameterized cross-layer mapping $\phi_{L-1} : \mathcal{M}_{L-1} \to \mathcal{M}_L$ as defined in Definition 2. This mapping induces a push-forward of measures such that $\mu_L^{\pm} = (\phi_{L-1})_* \mu_{L-1}^{\pm}$.

According to the change-of-variables formula in measure theory, for any measurable function g on $\mathcal{M}_L$, the integral with respect to the push-forward measure can be transformed into an integral over the original measure on $\mathcal{M}_{L-1}$:

$$\int_{\mathcal{M}_L} g(H^{(L)})\, d((\phi_{L-1})_* \mu_{L-1})(H^{(L)}) = \int_{\mathcal{M}_{L-1}} g(\phi_{L-1}(H^{(L-1)}))\, d\mu_{L-1}(H^{(L-1)}). \tag{136}$$

By setting $g = f_L^*$, the dual objective function at layer L can be expressed over the space $\mathcal{M}_{L-1}$:

$$\int_{\mathcal{M}_L} f_L^*\, d\mu_L^+ - \int_{\mathcal{M}_L} f_L^*\, d\mu_L^- = \int_{\mathcal{M}_{L-1}} (f_L^* \circ \phi_{L-1})\, d\mu_{L-1}^+ - \int_{\mathcal{M}_{L-1}} (f_L^* \circ \phi_{L-1})\, d\mu_{L-1}^-. \tag{137}$$

The composite function $f_L^* \circ \phi_{L-1}$ is precisely the pullback of f_L^* under ϕ_{L-1}, denoted by $\phi_{L-1}^* f_L^*$.

To verify the validity of this pullback function on $\mathcal{M}_{L-1}$, we examine its Lipschitz property. According to Assumption 2, the mapping ϕ_{L-1} is K_{L-1}-Lipschitz continuous, namely

$$d_{g^{(L)}}(\phi_{L-1}(H_x^{(L-1)}), \phi_{L-1}(H_y^{(L-1)})) \leq K_{L-1}\, d_{g^{(L-1)}}(H_x^{(L-1)}, H_y^{(L-1)}). \tag{138}$$

Combining this with the 1-Lipschitz property of f_L^*, we first have

$$|f_L^*(\phi_{L-1}(H_x^{(L-1)})) - f_L^*(\phi_{L-1}(H_y^{(L-1)}))| \leq d_{g^{(L)}}(\phi_{L-1}(H_x^{(L-1)}), \phi_{L-1}(H_y^{(L-1)})). \tag{139}$$

Then, by the K_{L-1}-Lipschitz continuity of ϕ_{L-1},

$$d_{g^{(L)}}(\phi_{L-1}(H_x^{(L-1)}), \phi_{L-1}(H_y^{(L-1)})) \le K_{L-1}\, d_{g^{(L-1)}}(H_x^{(L-1)}, H_y^{(L-1)}). \tag{140}$$

Therefore,

$$|f_L^*(\phi_{L-1}(H_x^{(L-1)})) - f_L^*(\phi_{L-1}(H_y^{(L-1)}))| \le K_{L-1}\, d_{g^{(L-1)}}(H_x^{(L-1)}, H_y^{(L-1)}). \tag{141}$$

Thus, the composite function $\phi_{L-1}^* f_L^*$ is K_{L-1}-Lipschitz on $\mathcal{M}_{L-1}$.

To construct a standard candidate scalar field satisfying the Kantorovich dual constraint, we introduce the normalization factor $\frac{1}{K_{L-1}}$ and define

$$f_{L-1}(H^{(L-1)}) = \frac{1}{K_{L-1}}(\phi_{L-1}^* f_L^*)(H^{(L-1)}) = \frac{1}{K_{L-1}} f_L^*(\phi_{L-1}(H^{(L-1)})). \tag{142}$$

It follows immediately that $\mathrm{Lip}(f_{L-1}) \le 1$. Therefore, the optimal information from layer L can be effectively transported back to layer $L-1$ via the pullback operator, inducing a valid candidate scalar field while avoiding repeated optimization of the full Kantorovich dual problem at every layer. □

D.2 Proof of Proposition: Cross-Layer Structural Suboptimality Bound

Proposition 2 (Cross-Layer Structural Suboptimality Bound). *Let $\phi_{L-1} : \mathcal{M}_{L-1} \to \mathcal{M}_L$, be the cross-layer mapping defined in Definition 2. The suboptimal error of the recursive candidate scalar field f_{L-1} is bounded by the layer-wise structural constants:*

$$\mathrm{Err}_{L-1} \le \Delta_{L-1}^* - \frac{\Delta_L^*}{\mathrm{Lip}(\sigma) \cdot s_1(W_{\text{in}}^{(L)}) \cdot \mathrm{Lip}(\mathrm{LN}) \cdot \big(1 + \mathrm{Lip}(\mathrm{Attn}_L)\,\mathrm{Lip}(\mathrm{LN})\big) \cdot s_1(W_{\text{out}}^{(L-1)})}, \tag{143}$$

where $\Delta_{L-1}^* = W_1(\mu_{L-1}^+, \mu_{L-1}^-)$ denotes the Wasserstein-1 separation at layer $L-1$, and $s_1(\cdot)$ denotes the spectral norm of the corresponding weight matrices. Here, $W_{\text{out}}^{(L-1)}$ and $W_{\text{in}}^{(L)}$ correspond to the output transformation of layer $L-1$ and the input transformation of layer L, respectively. Notably, the entire denominator constitutes the analytical upper bound of the cross-layer local Lipschitz constant K_{L-1}. Detailed explanations and proofs can be found in Appendix F.11.

Proof. According to the composition rule of Lipschitz mappings, the Lipschitz constant of $\phi_{L-1} = A_L \circ G_L \circ R_{L-1}$ satisfies $\mathrm{Lip}(\phi_{L-1}) \le \mathrm{Lip}(A_L)\,\mathrm{Lip}(G_L)\,\mathrm{Lip}(R_{L-1})$.

For the affine mapping R_{L-1}, its Lipschitz constant is determined by the operator norm of its weight matrix:

$$\mathrm{Lip}(R_{L-1}) = \|W_{\text{out}}^{(L-1)}\|_{\text{op}} = s_1(W_{\text{out}}^{(L-1)}). \tag{144}$$

For the intermediate update process G_L, applying the triangle inequality and the chain rule yields

$$\mathrm{Lip}(G_L) \le 1 + \mathrm{Lip}(\mathrm{Attn}_L)\,\mathrm{Lip}(\mathrm{LN}). \tag{145}$$

For the feed-forward projection A_L, similarly we obtain:

$$\mathrm{Lip}(A_L) \le \mathrm{Lip}(\sigma)\,\|W_{\text{in}}^{(L)}\|_{\text{op}}\,\mathrm{Lip}(\mathrm{LN}) = \mathrm{Lip}(\sigma)\,s_1(W_{\text{in}}^{(L)})\,\mathrm{Lip}(\mathrm{LN}). \tag{146}$$

Therefore, the explicit upper bound of K_{L-1} is

$$K_{L-1} := \mathrm{Lip}(\phi_{L-1}) \le \mathrm{Lip}(\sigma)\,s_1(W_{\text{in}}^{(L)})\,\mathrm{Lip}(\mathrm{LN})\Big(1 + \mathrm{Lip}(\mathrm{Attn}_L)\,\mathrm{Lip}(\mathrm{LN})\Big)\,s_1(W_{\text{out}}^{(L-1)}). \tag{147}$$

Since f_L^* is 1-Lipschitz, the composition rule gives

$$\mathrm{Lip}(f_L^* \circ \phi_{L-1}) \le \mathrm{Lip}(f_L^*)\,\mathrm{Lip}(\phi_{L-1}) \le 1 \cdot K_{L-1} = K_{L-1}. \tag{148}$$

Hence,

$$\mathrm{Lip}(f_{L-1}) = \mathrm{Lip}\left(\frac{1}{K_{L-1}} f_L^* \circ \phi_{L-1}\right) \le 1. \tag{149}$$

Thus, f_{L-1} is a valid candidate function for the Kantorovich dual problem at layer $L-1$.

According to the Kantorovich–Rubinstein duality theorem, the optimal separation distance at layer $L-1$ is

$$\Delta^*_{L-1} = \sup_{\mathrm{Lip}(f)\le 1} \left(\mathbb{E}_{\mu^+_{L-1}}[f] - \mathbb{E}_{\mu^-_{L-1}}[f] \right). \tag{150}$$

Since f_{L-1} is feasible, it follows that

$$\Delta^*_{L-1} \ge \mathbb{E}_{\mu^+_{L-1}}[f_{L-1}] - \mathbb{E}_{\mu^-_{L-1}}[f_{L-1}]. \tag{151}$$

Substituting the definition of f_{L-1} yields

$$\Delta^*_{L-1} \ge \frac{1}{K_{L-1}} \left(\mathbb{E}_{\mu^+_{L-1}}[f^*_L \circ \phi_{L-1}] - \mathbb{E}_{\mu^-_{L-1}}[f^*_L \circ \phi_{L-1}] \right). \tag{152}$$

By the change-of-variables theorem under the push-forward relation $\mu^\pm_L = (\phi_{L-1})_* \mu^\pm_{L-1}$, we have

$$\mathbb{E}_{\mu^\pm_{L-1}}[f^*_L \circ \phi_{L-1}] = \mathbb{E}_{\mu^\pm_L}[f^*_L]. \tag{153}$$

Therefore,

$$\Delta^*_{L-1} \ge \frac{1}{K_{L-1}} \left(\mathbb{E}_{\mu^+_L}[f^*_L] - \mathbb{E}_{\mu^-_L}[f^*_L] \right) = \frac{1}{K_{L-1}} \Delta^*_L. \tag{154}$$

The actual separation achieved by the recursive candidate scalar field is thus $\frac{1}{K_{L-1}} \Delta^*_L$. Define the suboptimal error as

$$\mathrm{Err}_{L-1} := \Delta^*_{L-1} - \left(\mathbb{E}_{\mu^+_{L-1}}[f_{L-1}] - \mathbb{E}_{\mu^-_{L-1}}[f_{L-1}] \right) = \Delta^*_{L-1} - \frac{1}{K_{L-1}} \Delta^*_L. \tag{155}$$

This quantity admits the standard interpretation of a suboptimality gap, measuring the discrepancy between the optimal achievable separation at layer $L-1$ and the separation attained by the recursively induced pullback candidate scalar field. Since we have already established a structural upper bound K^{bound}_{L-1} for K_{L-1}, it follows that

$$\frac{1}{K_{L-1}} \ge \frac{1}{K^{\mathrm{bound}}_{L-1}}, \tag{156}$$

and hence

$$-\frac{1}{K_{L-1}} \le -\frac{1}{K^{\mathrm{bound}}_{L-1}}. \tag{157}$$

Substituting this into the expression of the error yields the final explicit bound:

$$\mathrm{Err}_{L-1} \le \Delta^*_{L-1} - \frac{\Delta^*_L}{\mathrm{Lip}(\sigma)\, s_1(W^{(L)}_{\mathrm{in}})\, \mathrm{Lip}(\mathrm{LN}) \left(1 + \mathrm{Lip}(\mathrm{Attn}_L)\, \mathrm{Lip}(\mathrm{LN})\right) s_1(W^{(L-1)}_{\mathrm{out}})}. \tag{158}$$

This completes the proof. □

Algorithm 1 Dual Scalar Field Learning and Scalar Field-Guided Hallucination Mitigation

Require: Frozen LLM F_Θ, selected layer L; truthful dataset $\mathcal{D}^+$ and hallucinated dataset $\mathcal{D}^-$; scalar field network $\hat{f}_{\theta_L}$; intervention token ratio ρ; intervention strength α; Euler steps S; stability constant ϵ; Fisher regularization coefficient ξ.

Ensure: Empirical distributions μ_L^+, μ_L^-; learned scalar field f_L^*; mitigated activation $\tilde{H}_L$; recursive scalar field candidates $\mathcal{F} = \{f_\ell\}_{\ell=L-2}^{L}$.

Stage 1: Empirical Distribution Estimation and Scalar Field Training

Initialize activation sets: $\mathcal{A}_L^+ \leftarrow \emptyset, \quad \mathcal{A}_L^- \leftarrow \emptyset$
for each dataset $\mathcal{D}^s$, $s \in \{+, -\}$ **do**
  for each sample $(q, a^{(s)}) \in \mathcal{D}^s$ **do**
    Extract response token activations from layer L: $h_{L,t}^{(s)} = \mathrm{Act}_{L,t}(F_\Theta(q, a^{(s)}))$
    Update activation set: $\mathcal{A}_L^s \leftarrow \mathcal{A}_L^s \cup \{h_{L,t}^{(s)}\}_{t=1}^{T}$
  end for
  Construct empirical measure: $\mu_L^s(x) = \frac{1}{|\mathcal{A}_L^s|} \sum_{h \in \mathcal{A}_L^s} \delta(x - h)$
end for
Estimate the Fisher metric on the activation manifold:
$g_{\mu\nu}(h) = \mathbb{E}_{y \sim p_\Theta(\cdot|h)} \left[\frac{\partial \log p_\Theta(y|h)}{\partial h^\mu} \frac{\partial \log p_\Theta(y|h)}{\partial h^\nu} \right]$
Regularize the Fisher metric:
$g_{\mu\nu}(h) = g_{\mu\nu}(h) + \xi I_{\mu\nu}$
Compute the Lipschitz regularization over all collected activations:
$\mathcal{L}_{\mathrm{Lip}} = \frac{1}{2} \sum_{r \in \{+,-\}} \frac{1}{|\mathcal{A}_L^r|} \sum_{h \in \mathcal{A}_L^r} \left[\max\left(0, \left\| \nabla \hat{f}_{\theta_L}(h) \right\|_g - 1\right) \right]^2.$
Train scalar field $\hat{f}_{\theta_L}(h)$ via Kantorovich–Rubinstein duality:
$\theta_L^* = \arg\min_{\theta_L} \left(-\mathbb{E}_{h \sim \mu_L^+}[\hat{f}_{\theta_L}(h)] + \mathbb{E}_{h \sim \mu_L^-}[\hat{f}_{\theta_L}(h)] + \lambda_{\mathrm{Lip}} \mathcal{L}_{\mathrm{Lip}} \right)$
Obtain the learned scalar field: $f_L^* \leftarrow \hat{f}_{\theta_L^*}$

Stage 2: Scalar Field-Guided Hallucination Mitigation

Given a hallucination-prone activation H_L, compute the induced vector:
$v^\mu(H_L) = g^{\mu\nu}(H_L) \partial_\nu f_L^*(H_L)$
Select ρ proportion of generated tokens for intervention:
$\rho = 0.15$
Initialize mitigation trajectory: $H_L^{(0)} \leftarrow H_L$
for $s = 0$ to $S - 1$ **do**
  Update the flow direction:
  $v^\mu(H_L^{(s)}) = g^{\mu\nu}(H_L^{(s)}) \partial_\nu f_L^*(H_L^{(s)})$
  Compute the Riemannian norm:
  $\left\| v(H_L^{(s)}) \right\|_g = \sqrt{g_{\mu\nu}(H_L^{(s)}) v^\mu v^\nu}$
  Perform Euler integration:
  $H_L^{(s+1)} \leftarrow H_L^{(s)} + \alpha \frac{d(H_L^{(s)})}{\|v(H_L^{(s)})\|_g + \epsilon}$
end for
$\tilde{H}_L \leftarrow H_L^{(S)}$

Stage 3: Recursive Scalar Field Construction via Pullback Across Layers

Initialize recursive scalar field set: $\mathcal{F} \leftarrow \{f_L^*\}$
Initialize: $f_L \leftarrow f_L^*$
for $\ell = L$ to $L - 2$ **do**
  Obtain cross-layer mapping: $\phi_{\ell-1} : \mathcal{M}_{\ell-1} \to \mathcal{M}_\ell$
  Estimate the Lipschitz constant: $K_{\ell-1}$ of $\phi_{\ell-1}$
  Construct the lower layer scalar field by normalized pullback:
  $f_{\ell-1}(H^{(\ell-1)}) = \frac{1}{K_{\ell-1}} f_\ell \left(\phi_{\ell-1}(H^{(\ell-1)}) \right)$
  Update recursive scalar field set:
  $\mathcal{F} \leftarrow \mathcal{F} \cup \{f_{\ell-1}\}$
  Update recursive field: $f_\ell \leftarrow f_{\ell-1}$
end for
return $\mu_L^+, \mu_L^-, f_L^*, \tilde{H}_L, \mathcal{F}$

E EXPERIMENTAL SETUP

E.1 DATASETS AND METRICS.

We evaluate the proposed framework on two benchmark datasets. Our method was trained on a 20% sample dataset, and the remaining data was used for evaluation. The first is PDTB Prasad et al. (2008), for which performance is measured using the DISQ Score Miao et al. (2024). This metric evaluates discourse understanding from three complementary perspectives: Targeted, which reflects event-level accuracy; Counterfactual, which captures robustness; and Consistency, which measures logical coherence. These components are further combined into an Overall product score. The second benchmark is TruthfulQA Lin et al. (2022), which evaluates truthfulness through MC1 for single-answer accuracy, MC2 for multi-answer probability, Truthfulness for avoiding false statements, and Informativeness for response quality. Following the benchmark setting, we use true*informative as the composite evaluation metric.

E.2 BASELINES

We compared our proposed method with representative hallucination mitigation methods, including decoding-based methods CAD Shi et al. (2024), DoLa Chuang et al. (2024), and SH2 Kai et al. (2024), inference-time intervention methods ITI Li et al. (2023), and representation-editing methods TruthX Zhang et al. (2024), SPACE Wang et al. (2025a), ACT Wang et al. (2025b), TrFr Chen et al. (2024b), LITO Fatahi Bayat et al. (2024), SEA Qiu et al. (2024), SADI Wang et al. (2025c), HPR Pham & Nguyen (2024), and QRAE Liao et al. (2026). CAD enhances the model's reliance on external evidence by contrasting the output distributions under context-aware and context-free conditions. DoLa improves factuality by contrasting the logits from different transformer layers during decoding, encouraging the model to rely on earlier factual representations. SH2 further explores decoding-time truthfulness improvement by leveraging self-highlighted hesitation signals to guide the generation process toward more cautious and truthful responses. ITI shifts attention-head activations along a learned truthfulness direction to elicit more truthful answers without additional fine-tuning. Beyond inference-time strategies, representation-editing approaches directly modify internal model representations associated with truthfulness. TruthX edits internal model representations in a truthful latent space to alleviate hallucinations by steering hidden states toward truthful directions. SPACE extends representation editing by identifying and manipulating a shared activation subspace to jointly mitigate factuality and faithfulness hallucinations. TrFr introduces multi-scale truthfulness representation intervention, constructing hierarchical truth-related representations to improve the model's ability to distinguish truthful and misleading information. LITO learns task-specific intervention directions and uncertainty expressions to enhance truthfulness through adaptive representation modification. SEA performs spectral editing of activations, modifying the dominant components of hidden representations to improve model alignment and factual reliability. SADI further incorporates semantic-aware activation intervention by adapting editing operations according to the semantic properties of different inputs. HPR applies Householder pseudo-rotation to activation spaces from a direction-magnitude perspective, enabling more precise manipulation of truthfulness-related representations. QRAE introduces query-routed activation editing with truth-hierarchical preference optimization, dynamically selecting editing routes according to query characteristics and optimizing truthfulness preferences. ACT introduces Adaptive Activation Steering, a tuning-free activation editing method that improves truthfulness by steering hidden representations toward truthful directions during inference. Different from fixed steering approaches, ACT learns multiple truthfulness-related steering vectors for different hallucination categories and adaptively adjusts the steering intensity according to the current activation state.

E.3 IMPLEMENTATION DETAILS.

To evaluate the generalization ability across different model architectures, we first apply our method to several representative instruction-tuned LLMs, including DeepSeek-LLM-7B DeepSeek-AI et al. (2024), Qwen2-7B-Instruct Yang et al. (2024), LLaMA2-7B-Chat Touvron et al. (2023b), and LLaMA-3-Instruct-8B Grattafiori et al. (2024). The consistent improvements across these models demonstrate that our approach is model-agnostic and can effectively enhance truthfulness beyond a specific model family. Furthermore, to investigate the scalability with respect to model capacity, we conduct experiments on the LLaMA v1 family with different parameter sizes, including LLaMA-

Scalar Field Learning		Intervention and Pullback	
Learning rate	1×10^{-4}	ODE solver	First-order Euler
Weight decay	1×10^{-4}	Intervention intensity	30%
Batch size	128	Intervention token ratio	15%
Maximum epochs	8	Intervention layers	Last 3 layers
Dropout rate	10^{-1}	Euler steps	3
Lipschitz coefficient	10^{-1}	Pullback depth	2 layers
Fisher regularization coefficient	10^{-2}	Stability constant	10^{-4}

Table 2: Hyperparameter configurations for scalar field optimization and inference-time mitigation.

7B, LLaMA-13B, LLaMA-33B, and LLaMA-65B Touvron et al. (2023a). These results verify that our method consistently improves truthfulness across models with different scales. Finally, for the main comparison, we evaluate our approach on the Qwen2.5 Qwen et al. (2025) and LLaMA-3 Grattafiori et al. (2024) series and compare it with representative intervention methods under the same experimental settings. The results demonstrate that our method achieves superior performance over existing approaches in improving truthfulness. All models were loaded in FP32 format onto 8x NVIDIA A100 GPUs Choquette et al. (2021). To ensure fair and statistically robust comparisons, all methods were evaluated under the same inference settings, and the results were averaged from multiple independent runs to maintain consistency and reproducibility. We optimize the scalar field using AdamW Loshchilov & Hutter (2019) with a cosine decay learning-rate scheduler. The detailed hyperparameter configurations are provided in Table 2.

E.4 Practical Training of the Scalar Field.

In practice, directly optimizing over the entire Lipschitz function space in Eq. 4 is intractable for high-dimensional LLM activations. We therefore parameterize the optimal scalar field f_L^* using a lightweight two-layer MLP $\hat{f}_{\theta,L} : \mathbb{R}^d \to \mathbb{R}$. The nonlinear mapping of the MLP enables the scalar field to capture geometric structures in the activation space with low computational cost.

During training, the backbone LLM remains frozen, and only the parameters θ of $\hat{f}_{\theta,L}$ are optimized. Given a prompt response pair $(q, a^{(s)})$, where $s = 0$ and $s = 1$ denote hallucinated and truthful responses, respectively, the activations at layer L are extracted as $h_{L,t}^{(s)} = \mathrm{Act}_{L,t}\big(F_\Theta(q, a^{(s)})\big)$, $t = 1, \ldots, T$. Only response tokens are retained, while prompt and padding tokens are excluded. The collected activations from hallucinated and truthful responses are denoted as $\mathcal{A}_L^-$ and $\mathcal{A}_L^+$, respectively. These activation sets serve as empirical samples from the corresponding distributions μ_L^- and μ_L^+. The expectations in the Kantorovich–Rubinstein dual formulation are approximated by the empirical averages over these samples. Therefore, the scalar field is optimized with the objective:

$$\mathcal{L}_{\mathrm{SF}} = -\left[\frac{1}{|\mathcal{A}_L^+|}\sum_{h^+\in\mathcal{A}_L^+}\hat{f}_{\theta,L}(h^+) - \frac{1}{|\mathcal{A}_L^-|}\sum_{h^-\in\mathcal{A}_L^-}\hat{f}_{\theta,L}(h^-)\right] + \lambda_{\mathrm{Lip}}\mathcal{L}_{\mathrm{Lip}}. \tag{159}$$

The regularization term $\mathcal{L}_{\mathrm{Lip}}$ encourages the learned scalar field to satisfy the 1-Lipschitz constraint in Eq. 3 at the observed activation states. We evaluate the gradient norm at every truthful and hallucinated activation in $\mathcal{A}_L^+$ and $\mathcal{A}_L^-$, respectively. The Lipschitz regularization is defined as

$$\mathcal{L}_{\mathrm{Lip}} = \frac{1}{2}\sum_{r\in\{+,-\}}\frac{1}{|\mathcal{A}_L^r|}\sum_{h\in\mathcal{A}_L^r}\left[\max\left(0, \left\|\nabla\hat{f}_{\theta_L}(h)\right\|_g - 1\right)\right]^2. \tag{160}$$

Here, the gradient norm is the Riemannian norm defined in Definition 1. The regularization is evaluated pointwise over all collected truthful and hallucinated activations and penalizes the observed states whose gradient norm exceeds the Lipschitz bound. The detailed optimization procedure for training the scalar field is presented in Algorithm 1.

Method	True*Info↑	True↑	Info↑	MC1↑	MC2↑	MC3↑
Baselines						
Baseline	62.0	69.5	89.2	39.1	58.6	29.5
Supervised Fine-tuning	69.5	71.2	97.6	39.3	56.6	30.6
Few-shot Prompting	66.4	67.4	98.4	41.4	59.2	29.6
Decoding-based Methods						
DoLa	71.8	73.2	**98.0**	40.6	59.3	31.8
SH2	62.3	71.9	86.7	32.2	56.5	31.9
Editing-based Methods						
ITI	69.0	79.8	86.4	41.1	61.1	31.7
TrFr	73.2	82.0	89.3	41.5	60.0	30.8
TruthX	64.9	71.8	90.3	42.8	61.2	32.2
LITO	52.6	84.6	62.3	40.4	58.3	29.6
SEA	72.3	81.8	88.4	42.8	61.1	33.3
ACT	72.6	79.6	91.2	42.2	62.1	32.1
SADI	75.2	83.1	90.5	43.1	62.6	32.4
HPR	78.5	85.1	92.2	50.1	68.2	42.4
QRAE	85.1	89.6	95.0	63.3	77.9	55.0
Ours	**87.3**	**90.1**	96.5	**65.4**	**78.1**	**55.7**

Table 3: Comparison with existing methods on the TruthfulQA benchmark using LLaMA-3-Instruct-8B. The best results are highlighted in bold.

E.5 Practical ODE Intervention.

Although Eq. 11 formulates hallucination mitigation as a continuous first-order ODE, we use a fixed-step Euler discretization during inference to reduce computational overhead. Given the generated response, we compute the scalar field score for each token hidden state $H_{L,t}$ as $s_t = \hat{f}_{\theta,L}(H_{L,t})$. The tokens with the lowest ρ proportion of scores are selected for intervention: $\mathcal{I} = \mathrm{TopK}_t(-s_t, \rho T)$, where $\rho = 0.15$. The intervention is applied to the selected tokens across the final three Transformer layers. The scalar field at the final layer is directly optimized, while the scalar fields for the preceding two layers are obtained through the pullback mechanism.

Given the learned scalar field $\hat{f}_{\theta,L}$, we compute the scalar field-induced tangent component

$$v^{\mu}(H_L) = g^{\mu\nu}(H_L)\frac{\partial \hat{f}_{\theta,L}(H_L)}{\partial H_L^{\nu}}. \tag{161}$$

For the selected tokens, their hidden states are updated along the scalar field-guided direction using a fixed-step Euler discretization:

$$H_L^{(s+1)} = H_L^{(s)} + \alpha \cdot \frac{v(H_L^{(s)})}{\|v(H_L^{(s)})\|_g + \epsilon}, \quad s = 0, \ldots, S-1, \tag{162}$$

where α is the intervention strength set to 0.3, S is the number of Euler steps set to 3, and ϵ is a small constant for numerical stability. The detailed hallucination mitigation are presented in Algorithm 1.

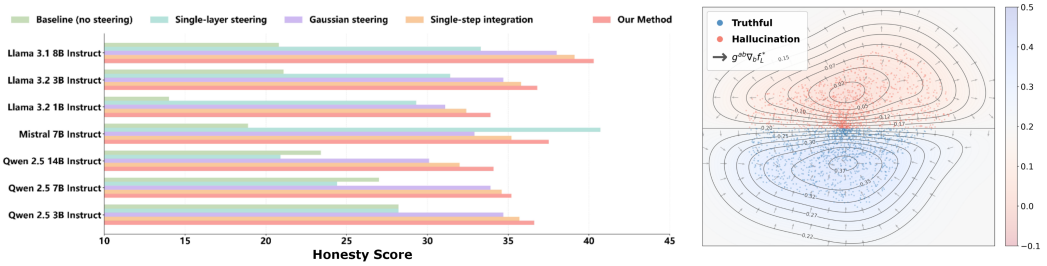
F Experimental Validation

F.1 Cross-Model Comparison with Baselines

We compare our method with representative baselines on the TruthfulQA benchmark using LLaMA-3-Instruct-8B. As shown in Table 3, our method achieves the best overall performance across most

Model	Method	BLEURT	TRUE	True×Info	MC1	MC2	Avg.
LLaMA-7B	Base	49.1	43.2	39.5	35.1	50.7	43.5
	ACT	57.3	54.2	46.6	35.5	52.3	49.2
	Ours	**60.7**	**57.1**	**59.5**	**36.1**	**53.4**	**53.4**
LLaMA-13B	Base	59.7	51.3	43.4	39.1	55.1	49.7
	ACT	69.6	67.0	46.0	41.4	59.1	56.6
	Ours	**70.8**	**69.3**	**48.3**	**44.2**	**61.8**	**58.9**
LLaMA-33B	Base	62.9	52.0	42.8	41.9	58.6	51.6
	ACT	71.9	65.2	49.6	44.2	62.3	58.6
	Ours	**73.5**	**67.9**	**53.6**	**45.9**	**65.1**	**61.2**
LLaMA-65B	Base	68.8	58.1	48.8	45.5	62.9	56.8
	ACT	76.1	72.3	50.4	46.3	64.7	61.9
	Ours	**79.6**	**76.2**	**53.1**	**49.5**	**66.4**	**65.0**

Table 4: Truthfulness evaluation on the TruthfulQA benchmark with different LLaMA backbones. Our method is compared with ACT Wang et al. (2025b). The best results are highlighted in bold.



(a) Comparison of different steering methods

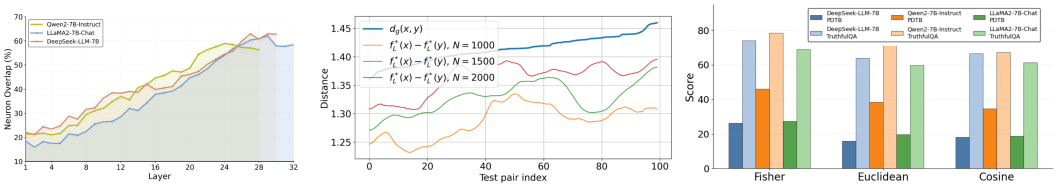
(b) Visualization of scalar field tangent vectors and isocontours

Figure 5: (a) compares the truthfulness improvement performance of different steering strategies across various language models. The results include baseline models without intervention, single-layer steering, Gaussian steering, single-step integration based on the initial tangent vector, and our proposed method. (b) presents the PCA-based visualization of the tangent directions of the learned scalar field and its isocontours after normalizing selected activations onto a hypersphere.

truthfulness metrics. Specifically, our approach obtains 87.3 on True*Info and 90.1 on True, outperforming previous approaches and demonstrating its effectiveness in improving factual reliability. Compared with decoding-based methods, our method achieves consistent improvements without relying on inference-time decoding heuristics. Compared with existing representation editing methods, our approach provides substantial gains on multiple-choice truthfulness metrics, improving MC1, MC2, and MC3 to 65.4, 78.1, and 55.7, respectively. These improvements indicate that the proposed scalar field-guided intervention more effectively reshapes internal representations toward truthful generation. Notably, our method achieves competitive performance on Info while obtaining the highest truthfulness scores, suggesting that the intervention improves factual accuracy without sacrificing response informativeness. To further investigate the scalability and robustness of our approach across different model capacities, we conduct additional experiments on LLaMA models with various parameter sizes, as reported in Table 4. Our method consistently improves the baseline models across all scales, including LLaMA-7B, LLaMA-13B, LLaMA-33B, and LLaMA-65B. For example, on LLaMA-65B, our method improves the BLEURT score from 68.8 to 79.6, TRUE from 58.1 to 76.2, and MC2 from 62.9 to 66.4, demonstrating that the proposed intervention remains effective even with stronger base models. Compared with ACT, our approach achieves further gains across almost all evaluation metrics, with the largest improvement observed on the True×Info metric. These results verify that our scalar field-guided intervention is model-size ag-

Variant	DeepSeek-LLM-7B		Qwen2-7B-Instruct		LLaMA2-7B-Chat	
	PDTB	TruthfulQA	PDTB	TruthfulQA	PDTB	TruthfulQA
Direct multi-layer fields	**26.8**	**74.3**	**46.2**	**79.1**	**28.2**	**69.5**
Full: 3-layer Pullback	26.4	74.1	46.2	78.4	27.5	69.0
w/o Lipschitz	25.6	73.4	45.5	77.8	26.2	67.5
One-step shift	23.2	72.4	42.1	75.8	26.6	65.2
Naive cross-layer copy	18.6	68.7	37.1	71.7	22.6	60.7
w/o Pullback	25.3	73.3	44.8	77.8	26.2	68.0
Random-neuron intervention	17.3	63.7	37.6	68.6	16.8	60.7
w/o Fisher Metric	16.0	64.0	38.5	71.5	19.6	59.7

Table 5: Ablation study of different components in our framework across three backbone models. Direct multi-layer fields independently optimize the scalar field at each Transformer layer. Full: 3-layer Pullback denotes scalar fields are optimized at last layers and recursively propagated across three consecutive layers. w/o Lipschitz removes the Lipschitz regularization term in Eq. 159 during scalar field optimization. Naive cross-layer copy replaces the normalized pullback operator with direct scalar field copying across layers. w/o Pullback removes the recursive Pullback mechanism. Random-neuron intervention replaces geometry-guided neuron selection with randomly selected neurons. w/o Fisher Metric replaces the Fisher metric with the Euclidean metric.



(a) Neuron overlap between probes and scalar fields

(b) Saturation behavior with increasing sample

(c) Comparison across different metrics.

Figure 6: (a) Neuron overlap ratio between probe-based and scalar field-based identification across layers. (b) Saturation behavior of scalar field distances with increasing sample size, where $d_g(x, y)$ is compared with the learned metric $f_L^*(x) - f_L^*(y)$ under different sample sizes. (c) Comparison of different distance metrics (Fisher, Euclidean, and Cosine) across multiple language models and evaluation datasets.

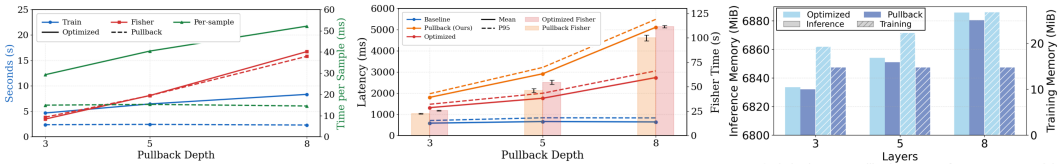
nostic and can consistently enhance truthfulness by modifying internal representations rather than relying on model-specific decoding strategies.

F.2 Ablation Experiment

We conduct a comprehensive ablation study across three backbone models and two benchmarks, as shown in Table 5. The results demonstrate that the proposed recursive pullback mechanism effectively approximates independently optimized multi-layer scalar fields. Specifically, the Full model achieves performance comparable to Direct multi-layer fields while requiring only a single directly trained scalar field, validating the effectiveness of the proposed cross-layer field propagation. Removing the pullback mechanism leads to consistent performance degradation across all models, indicating that directly applying the final-layer scalar field is insufficient for modeling layer-wise hallucination patterns. Moreover, naive cross-layer copying causes a substantial drop, highlighting the importance of the proposed geometric pullback transformation rather than simple parameter sharing. The Lipschitz regularization also provides consistent improvements, demonstrating that constraining the scalar field smoothness helps produce more stable intervention trajectories. Replacing the iterative Euler-based correction with a one-step shift significantly decreases performance, confirming the necessity of continuously updating the intervention direction along the mitigation trajectory. Finally, random neuron intervention results in severe degradation, validating the effectiveness of H-Neuron selection. Replacing the Fisher information metric with the Euclidean metric also causes a notable performance drop, confirming that the underlying activation geometry plays an important role in identifying effective correction directions.

Model	α(%)	TruthfulQA				
		True*Info(%)	MC1(%)	MC2(%)	True(%)	Info(%)
Qwen2-7B	10	0.7	1.1	0.5	0.6	0.3
	20	2.5	2.1	1.2	1.8	1.6
	30	3.4	3.8	1.4	1.9	1.8
	40	1.9	2.4	0.7	0.5	0.9
	50	-1.3	-1.4	-0.2	-1.1	-1.5
Llama2-7B	10	2.0	2.7	3.1	1.4	0.6
	20	2.6	3.2	3.9	0.9	0.3
	30	2.8	3.4	4.5	1.5	0.6
	40	-0.5	-1.1	-2.5	-0.2	-0.1
	50	-1.9	-2.4	-3.2	-1.6	-1.8
DeepSeek-LLM-7B	10	1.4	5.1	1.0	1.1	0.2
	20	2.9	8.3	2.6	2.1	0.7
	30	3.1	9.3	2.8	3.9	0.9
	40	1.2	3.5	2.2	0.6	0.3
	50	-1.7	-1.7	-2.6	-1.2	-1.6

Table 6: Intervention strength analysis for scalar field-guided H-Neuron repair on TruthfulQA. The intervention strength α controls how strongly hallucination-associated activations are shifted toward the truthful direction. All values denote the magnitude of score changes in percentage points.



(a) Training-time computational cost comparison

(b) Inference-time computational cost comparison

(c) GPU memory usage comparison

Figure 7: Efficiency and memory analysis of the Optimized and Pullback methods on Qwen2.5-1.5B-Instruct. (a) Training efficiency comparison with different pullback depths, including optimization cost, online Fisher computation time, and per-sample overhead measured on 160 TruthfulQA samples. (b) Inference efficiency comparison on 60 TruthfulQA samples, reporting response latency, P95 latency, and online Fisher computation cost. (c) GPU memory consumption comparison under different numbers of covered layers. Solid and hatched bars denote inference-time peak memory and training memory usage, respectively.

F.3 Effect of Intervention Strength

Table 6 further analyzes the effect of different intervention strengths on scalar field-guided H-Neuron repair. The results show that when the intervention intensity was too low, it was insufficient to effectively transfer hallucination-related activation toward the truthful region. In contrast, a moderate intervention strength produces more stable and consistent improvements across different models and metrics, suggesting that an appropriate correction magnitude can suppress hallucination-related activation components while largely preserving the original semantic structure of the hidden representations. However, when the intervention becomes too strong, the repair effect begins to deteriorate, indicating that excessive modification may over-correct the activation state and disrupt useful semantic information. These observations demonstrate that scalar field-guided repair should be understood as a controlled directional correction along the truthfulness-oriented scalar field, rather than as aggressive activation suppression.

F.4 Computational Cost of Pullback versus Scalar Field Optimization

We analyze the computational efficiency and GPU memory consumption of the optimized and pullback scalar fields under different pullback depths. Experiments are conducted on Qwen2.5-1.5B-Instruct using 160 TruthfulQA samples for optimization and 60 samples for inference evaluation. As

shown in Figure 7, the proposed pullback mechanism substantially reduces optimization overhead while maintaining comparable inference efficiency and memory consumption. During training-time optimization, the pullback scalar field achieves significant computational savings compared with the optimized scalar field. As the pullback depth increases, the optimization cost of the optimized field grows rapidly, whereas the pullback field maintains a relatively stable training overhead. This indicates that the pullback mechanism avoids repeatedly optimizing high-dimensional scalar fields and provides a more efficient optimization process. Moreover, the GPU memory consumption during training remains nearly constant for the pullback field, while the optimized field requires increasingly larger memory as more layers are involved, demonstrating the improved memory efficiency of the proposed approach. During inference, the pullback mechanism introduces additional online computation due to the required field transformation, resulting in increased response latency and reduced token throughput. However, this overhead remains limited compared with the overall inference process. The peak GPU memory consumption increases with the number of covered layers for both methods, reflecting the additional computation involved in deeper field transformations. Notably, the pullback method consistently achieves slightly lower peak GPU memory usage than the optimized field under different layer configurations, indicating that the proposed approach can reduce memory requirements while preserving comparable inference performance. Furthermore, we evaluate the computational overhead of online Fisher matrix estimation. The results show that the pullback mechanism reduces Fisher computation time compared with the optimized scalar field, suggesting that the proposed method not only lowers the main optimization cost but also alleviates the additional overhead introduced by online Fisher estimation.

F.5 Comparison with Existing Activation Steering Baselines

Figure 5(a) compares different activation steering strategies on seven instruction-tuned language models. Single-layer steering intervenes at a selected Transformer layer by extracting an truthfulness steering direction from the representation differences between honest and dishonest samples and injecting it into the hidden state. While effective, it assumes that truthfulness-related representations are concentrated in specific layers, which may limit its robustness across models. Gaussian steering extends this strategy by distributing the steering signal across consecutive layers with Gaussian weighting. This smooth cross-layer intervention assumes that high-level attributes such as truthfulness emerge progressively across layers, reducing potential representation disruption caused by single-layer modification. As shown, Gaussian steering generally achieves more stable improvements than single-layer steering. The single-step integration method estimates the steering trajectory from the initial tangent vector through exponential mapping, i.e., $\exp(g^{ab}\nabla_b f_L^*(\gamma(t)))$, providing a geometry-inspired approximation of the optimal intervention direction while better preserving the representation structure. Our method achieves the strongest overall performance across most models, demonstrating that scalar field-guided intervention can more accurately identify and exploit truthfulness-related directions in the representation space. These results highlight the advantage of geometry-aware steering over layer-specific heuristic approaches.

F.6 Geometric Visualization of the Scalar Field

Figure 5(b) illustrates the PCA projection of the learned scalar field on the TruthfulQA dataset. We first normalize the model activations onto a hypersphere and train a scalar field over this activation manifold. The PCA visualization shows the tangent vector directions induced by the scalar field at different points, along with the associated isocontours. The induced tangent vectors generally point toward the truthful region, suggesting that the learned scalar field captures a meaningful steering direction. Meanwhile, the scalar values smoothly decrease from the truthful region toward the hallucination region, demonstrating a structured honesty landscape encoded by the scalar field.

F.7 Consistency between Probe and Scalar field Neuron Identification

To evaluate the reliability of the scalar field neuron identification method, we compare the H-Neurons identified by the scalar field with those obtained by the probing method. As shown in Figure 6(a), we calculate the overlap ratio between the neuron sets identified by the two approaches across different layers. The results show that the neuron overlap ratio between the two methods generally increases as the model depth increases, indicating that the scalar field-based method captures hallucination-related neuron patterns that are increasingly consistent with those identified by

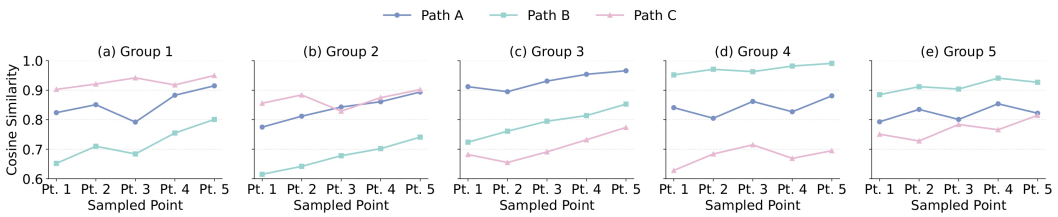


Figure 8: Cosine similarity between the scalar field-induced direction $g^{ab}\nabla_b f_L^*(\bar{\gamma}(t))$ and the normalized trajectory tangent $\dot{\bar{\gamma}}^a(t)/\|\dot{\bar{\gamma}}(t)\|_g$ on the TruthfulQA benchmark, where $\bar{\gamma}(t) = \gamma(t)/\|\gamma(t)\|_2$ denotes the activation state normalized onto the unit hypersphere in the original high-dimensional activation space. Five sample groups are randomly selected, each containing 100 response samples and three mitigation trajectories. The similarity is evaluated at five sampled points along each trajectory.

probing in deeper layers. The relatively lower overlap in shallow layers may be attributed to the approximation errors introduced by the scalar field pullback process. Meanwhile, the existence of partially non-overlapping neurons suggests that the scalar field approach may provide additional capability for discovering distinctive neuron features beyond those captured by conventional probing methods.

F.8 Saturation Behavior under Increasing Sampling Size

We further investigate the saturation property of the proposed scalar field metric with respect to the number of sampled data points. Figure 6(b) presents the comparison between the geometric distance $d_g(x, y)$ and the learned scalar field difference $f_L^*(x) - f_L^*(y)$ under different sampling sizes. As the sample size increases from $N = 1000$ to $N = 2000$, the learned distances gradually stabilize, exhibiting a clear saturation trend. Meanwhile, all learned scalar field distances remain bounded below the reference metric $d_g(x, y)$, which is consistent with the theoretical property induced by the Lipschitz constraint. The convergence behavior indicates that increasing the number of samples facilitates satisfaction of the dual saturation condition.

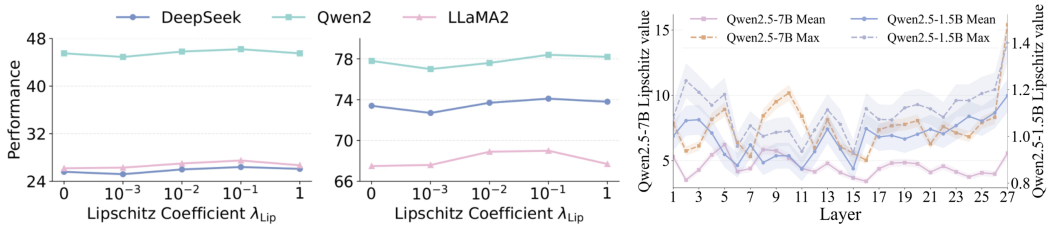
F.9 Comparison of Different Distance Metrics

We compare the proposed metric with commonly used distance measures, including Fisher, Euclidean, and Cosine distances. Figure 6(c) reports the performance of different metrics on multiple language models and evaluation datasets. Across different models, the Fisher-based metric achieves competitive performance, while Euclidean and Cosine metrics exhibit different degrees of degradation depending on the model and dataset. These results suggest that the Fisher metric is a more suitable choice for characterizing activation differences, as it incorporates the underlying geometric structure of the model distribution and provides a more effective measure for capturing behavior-related variations in language model activation.

F.10 Trajectory Consistency of Scalar Field-Guided Mitigation

Theorem 1 establishes that the scalar field-induced direction $g^{ab}\nabla_b f_L^*(\gamma(t))$ coincides almost everywhere with the normalized geodesic tangent $\frac{\dot{\gamma}^a(t)}{\|\dot{\gamma}(t)\|_g}$ along geodesics connecting optimal transport pairs in $\mathrm{supp}(\pi^*)$ that satisfy the duality saturation condition. This equivalence is therefore theoretically guaranteed only along the supported optimal transport geodesics.

For the empirical evaluation, all activation states along a mitigation trajectory are normalized onto the unit hypersphere in the original high-dimensional activation space. We further examine whether the directional consistency predicted by Theorem 1 can also be observed along general mitigation trajectories, which are not explicitly assumed to belong to the support of the optimal transport plan. As shown in Figure 8, we randomly sample five groups from the TruthfulQA benchmark. Each group contains 100 response samples and three independent mitigation trajectories, the cosine similarity is evaluated at five evenly sampled points on the normalized high-dimensional hypersphere. The cosine similarity remains consistently high across the sampled groups and trajectories. Although



(a) LLM Performance across Varying lipschitz coefficient

(b) Trends of lipschitz empirical values across layers

Figure 9: (a) Sensitivity analysis of the Lipschitz coefficient λ_{Lip} on PDTB (left) and TruthfulQA (right). We evaluate different values of λ_{Lip} using three backbone models, including DeepSeek, Qwen2, and LLaMA2. The results show that the proposed scalar field remains stable across a wide range of Lipschitz constraints, demonstrating the robustness of the learned geometric representation. (b) The distribution of Lipschitz values across different network layers for Qwen2.5-7B and Qwen2.5-1.5B. The solid lines represent the mean value for each layer, while the dashed lines indicate the maximum value.

the exact equivalence in Theorem 1 is restricted to supported optimal transport geodesics, strong directional alignment is also empirically observed along more general mitigation trajectories on the normalized high-dimensional hypersphere. This observation suggests that the learned scalar field provides a stable local mitigation direction beyond the idealized setting covered by the theorem.

F.11 Empirical Verification of Lipschitz Constants

As illustrated in Figure 9(b), we first present the empirical variation of the cross-layer Lipschitz constants across different network layers. This empirical observation is fundamentally tied to the theoretical error bounds of our recursive scalar field construction (Proposition 1), which avoids redundant optimization of the Kantorovich dual objective at every layer. To theoretically justify the effectiveness of this mechanism, Proposition 2 establishes an explicit upper bound on the approximation error that directly depends on these layer-wise structural constants. Crucially, as established in the proof of Proposition 2, the suboptimality gap is explicitly given by $\mathrm{Err}_{L-1} = \Delta^*_{L-1} - \frac{1}{K_{L-1}}\Delta^*_L$, where the subtracted term $\frac{1}{K_{L-1}}\Delta^*_L$ represents the actual separation distance achieved by our recursively induced candidate. The denominator K_{L-1} is structurally determined by the composition of layer-wise Lipschitz constants, including the spectral norms of weight matrices and the properties of activation/normalization functions. If K_{L-1} were to grow uncontrollably large, the subtracted term would vanish, rendering the error bound vacuous. However, our empirical statistics indicate that K_{L-1} does not grow excessively large, allowing the recursive pullback separation term $\frac{1}{K_{L-1}}\Delta^*_L$ to remain effective. Consequently, the theoretical upper bound on the approximation error Err_{L-1} is kept within a reasonable range. This consistency between our empirical layer-wise analysis and the analytical suboptimality bound suggests that the recursive pullback mechanism introduces an acceptable level of approximation error, validating the efficacy of the proposed geometric representation without requiring redundant per-layer optimization.

F.12 Hyperparameter Experimentation

We further investigate the effect of the Lipschitz coefficient λ_{Lip} (Eq. 159) on the performance of our method. As shown in Figure 9(a), the performance remains relatively stable across different values of λ_{Lip} on both PDTB and TruthfulQA. This indicates that the Lipschitz regularization provides a smooth constraint on the learned scalar field without requiring sensitive hyperparameter tuning. Extremely small or large coefficients do not lead to significant degradation, demonstrating the robustness of the proposed geometric formulation.

G LIMITATIONS

Although our framework provides a geometrically principled interpretation for H-Neuron detection and mitigation, several theoretical components rely on idealized assumptions. First, the activation space is modeled as a locally smooth Riemannian manifold, and the truthful and hallucinated activations are treated as empirical distributions on this manifold. This assumption provides a useful geometric abstraction, but it does not imply that the full high-dimensional activation space of an LLM globally satisfies a smooth manifold structure. Second, the equivalence between the raised gradient of the optimal scalar field and the normalized geodesic tangent direction is derived under the saturation condition of Kantorovich–Rubinstein duality along optimal transport paths. In practice, however, the scalar field is approximated by a parameterized neural network trained from finite activation samples, and therefore the learned field should be regarded as an empirical approximation to the ideal Kantorovich potential rather than an exact optimizer over the full Lipschitz function space. Third, the proposed mitigation dynamics are implemented through a discretized local update rather than by solving the continuous ODE exactly. Consequently, our intervention should be interpreted as a local scalar field-guided correction rather than a globally optimal transport trajectory. These limitations do not invalidate the proposed framework, but they indicate that the theoretical results serve as principled guidance for designing and interpreting the intervention mechanism, while the practical effectiveness still needs to be verified empirically across models, datasets, and intervention settings.