Detection without Expression: A Geometric perspective of Language Model Hallucination

Anonymous authors
Paper under double-blind review

Abstract

Language models often respond fluently and confidently to questions for which the appropriate response would be to abstain. We study cases where the prompt is underspecified, has a false premise, or is outside the model's reliable knowledge. Such errors are usually treated as failures of factual access. We argue that they also reflect a failure of routing. A model may internally represent that an input should not be answered while failing to transform that representation into output behavior. Cross-entropy training creates prediction-aligned directions through which token commitments are expressed, because each example supplies a sharp gradient toward a vocabulary target. Answerability, however, is not given an equally stable target unless the training distribution explicitly rewards abstention. It can therefore be encoded as an input-aligned feature of the residual stream without becoming a prediction-aligned control variable. In this view, hallucination can be understood as a mismatch between the geometry that detects uncertainty and the geometry that expresses decisions. Across autoregressive transformer families, we find that factual and uncertain prompts are strongly separated in hidden states, while standard output-side uncertainty measures expose only a weak trace of this distinction. The answerability boundary is concentrated in the principal input geometry and only inconsistently aligned with the prediction geometry defined by the unembedding. Causal interventions confirm that this geometry is not merely diagnostic: routing the hidden answerability signal directly to refusal logits produces selective abstention, boundary steering produces large direction-dependent shifts in decoded responses, and linear projection onto the factual subspace does not repair uncertain states. These results suggest that reducing hallucination requires mechanisms that explicitly connect internal answerability representations to the output pathways where linguistic commitments are made.

1 Introduction

Large language models can produce fluent answers to questions for which the correct response is to abstain, ask for clarification, or state that the answer is not known. The standard explanation is a knowledge deficit: the model lacks the relevant fact, so it fabricates one. This view motivates retrieval augmentation, knowledge editing, supervised fine-tuning, and other methods that attempt to provide missing information or reward factual behavior (Lewis et al., 2020; Kandpal et al., 2023; Lv et al., 2024). These methods are important, but they do not fully explain why models often answer false-presupposition or underspecified prompts with confidence. In those cases, hallucination reflects a decision not to express uncertainty.

This paper studies the geometry of that decision. We ask whether transformer language models internally distinguish answerable from unanswerable inputs, and whether that distinction is represented in directions that can affect the final token distribution. The answer is split. The distinction is highly visible in intermediate residual-stream activations. A linear probe trained on hidden states separates factual from uncertain prompts with high accuracy across model families. Yet simple output-side uncertainty measures, such as entropy and maximum probability, are poor detectors. The model often contains a usable answerability signal, but the signal does not behave like a normal prediction variable.

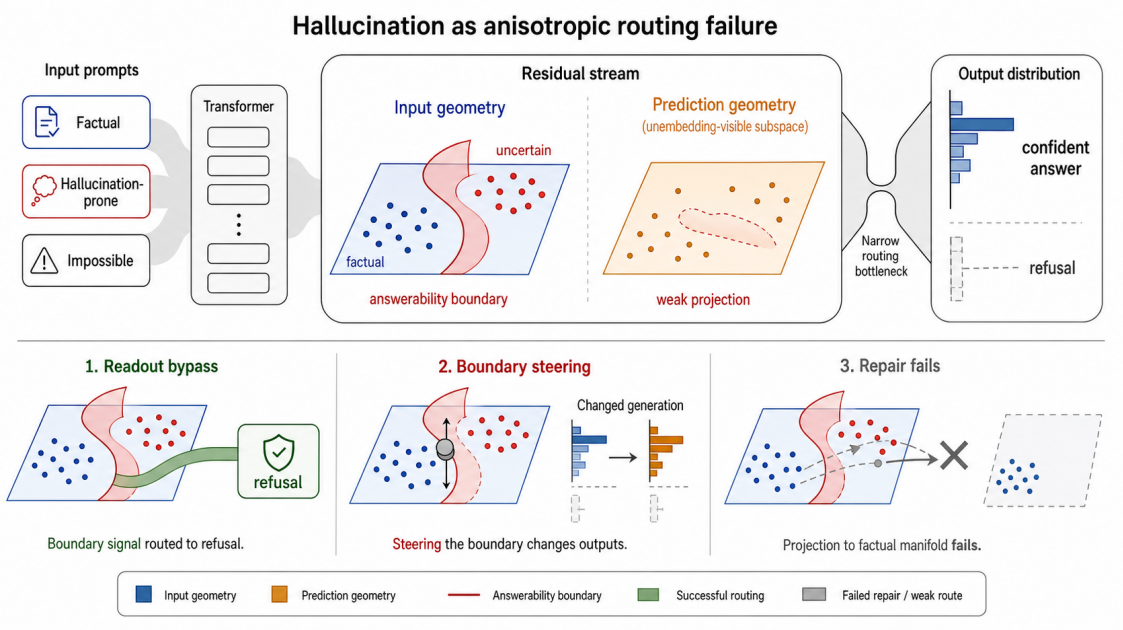


Figure 1: Answerability is represented in the residual stream as an input-aligned boundary separating factual and uncertain prompts. Because this direction is only weakly and inconsistently aligned with prediction geometry, the signal is not reliably expressed in the output distribution. A readout bypass restores expression by directly routing the hidden signal to refusal logits; steering confirms the boundary is causal; linear manifold repair fails because uncertain representations are not simple displaced factual states.

We explain this dissociation using an anisotropic view of residual-stream geometry. The unembedding matrix defines a prediction-aligned subspace: directions in the residual stream that strongly affect logits. Separately, the distribution of activations defines an input-aligned subspace: directions that carry the principal variation of the residual stream as prompts are processed. Nothing in the training objective requires these two geometries to align. Cross-entropy training provides sharp supervision to token-prediction directions because each example pulls probability mass toward a vocabulary vertex. It does not, by itself, give a stable target for abstention on underspecified or false-premise inputs. Therefore, a model can learn to detect answerability as an input-side feature while failing to transform it into a prediction-side control variable.

The empirical picture is consistent with this hypothesis: across the models tested, the answerability boundary is concentrated in the input manifold: its relative subspace fraction in the principal input subspace is approximately one. Its projection into the prediction subspace is smaller, sometimes moderate, and sometimes nearly zero. Thus the claim is not that uncertainty is always absent from the unembedding subspace, but that answerability is routed differently from ordinary token prediction. It's a residual-stream feature with strong input geometry and weak or unreliable output geometry.

We make four contributions: first, we develop a residual-stream anisotropy framework for hallucination that separates internal detection from output expression. Second, we define measurements of answerability geometry using boundary directions, input and output projectors, relative subspace fraction, intrinsic dimensionality, and persistent homology. Third, we report comprehensive experiments on autoregressive transformers showing that hidden-state probes detect uncertain prompts much better than output confidence or entropy. Fourth, we use three causal interventions, readout bypass, boundary steering, and manifold repair, to show that the signal is usable, causally relevant, and not recoverable by a simple linear projection onto factual representations.

2 Anisotropic Routing of Answerability

A decoder-only transformer maintains a residual stream $\mathbf{h}^{(\ell)} \in \mathbb{R}^d$. Each layer updates this stream through attention and MLP sublayers,

$$\mathbf{h}^{(\ell+1)} = \mathbf{h}^{(\ell)} + \mathrm{Attn}_\ell(\mathbf{h}^{(\ell)}) + \mathrm{MLP}_\ell(\mathbf{h}^{(\ell)}). \tag{1}$$

The final residual stream is mapped to logits by the unembedding matrix $W_U \in \mathbb{R}^{V \times d}$:

$$\mathbf{z} = W_U \mathbf{h}^{(L)}, \qquad \mathbf{p} = \mathrm{softmax}(\mathbf{z}). \tag{2}$$

The right singular vectors of W_U define directions in residual-stream space that have direct access to the vocabulary. We call the span of the top k such directions the prediction subspace and denote its projector by Π_{pred}. The residual-stream activations themselves define another geometry. The top k principal components of hidden states at a given layer define an input subspace, with projector Π_{input}. The first is set by the output map. The second is set by the distribution of representations.

These two geometries are shaped by different sources of anisotropy. For a linear map W, the per-example gradient has outer-product form

$$\nabla_W \mathcal{L} = \boldsymbol{\delta} \mathbf{a}^\top. \tag{3}$$

where $\mathbf{a}$ is the activation entering the map and $\boldsymbol{\delta}$ is the upstream gradient. A matrix that reads from the residual stream inherits structure from the covariance of $\mathbf{h}^{(\ell)}$. A matrix that writes to the residual stream inherits structure from upstream gradients. Under cross-entropy, the logit gradient is

$$\nabla_{\mathbf{z}} \mathcal{L} = \mathbf{p} - \mathbf{e}_{y^*}, \tag{4}$$

so each example supplies a sharp direction associated with a target vocabulary vertex. This is the simplex-vertex attractor: training repeatedly moves probability mass toward one-hot targets. After multiplication by $W_U^\top$, these target-token gradients enter the residual stream through prediction geometry.

Answerability is different from ordinary next-token commitment: a prompt can be underspecified or have a false-premise, but ordinary pretraining still presents a next token rather than a special abstention target. Unless the training distribution contains a consistent target for refusal or uncertainty, the loss does not require the answerability feature to become a prediction-aligned direction. It may remain an input-aligned latent variable: useful for internal processing, linearly accessible to a probe, but weakly coupled to the logits that express behavior.

This framework suggests a specific pattern: first, factual and uncertain prompts should be separable in hidden states. Second, output confidence and entropy should be worse uncertainty detectors than hidden-state probes. Third, the answerability boundary should have high overlap with Π_{input} and weaker or less stable overlap with Π_{pred}. Fourth, directly connecting the hidden-state signal to refusal logits should induce abstention. Fifth, finite steering along the boundary should change token generation, even if small local perturbations have a weak output effect. Finally, if uncertain representations fragment into disconnected regions rather than forming one abstention state, projecting them onto the factual manifold should not repair generation.

3 Methodology

3.1 Models and Data

We evaluate autoregressive transformers from the Llama, Mistral, and Qwen families. The main geometric experiments include Llama-3.1-8B, Llama-3.2-3B, Llama-3.2-3B-IT, Mistral-7B-v0.1, Qwen-2.5-7B, Qwen-2.5-3B-IT, Qwen3-8B, and Qwen3-32B. The comprehensive causal sweeps use the models for which all intervention runs were completed: Llama-3.1-8B, Llama-3.2-3B, Llama-3.2-3B-IT, Mistral-7B-v0.1, Qwen-2.5-7B, Qwen-2.5-3B-IT, Qwen3-8B, and Qwen3-32B.

We use three language-model prompt splits. The *Factual* split contains manually verified questions on which target models answer correctly at high rates. The *Impossible* split contains underspecified prompts and

false-presupposition prompts for which an appropriate model should abstain or ask for clarification. The *Hallucination* split combines TruthfulQA-style questions and long-tail PopQA-style questions on which models are prone to confident error. The splits are designed to separate two notions: factual recall and answerability. Factual questions are answerable; impossible prompts are not answerable as stated; hallucination prompts are answerable in principle but unreliable for the model.

We also test whether the geometry could be explained by superficial dataset differences. For every question we compute token count, dependency-tree depth, clause count, Flesch–Kincaid grade level, and reading ease. The factual and impossible splits are closely matched on these measures, and the impossible split is marginally simpler on several of them. The hallucination split is slightly longer, as expected for long-tail factual questions, but the central factual-versus-impossible contrast is not explained by syntactic complexity. Full statistics and details about the datset are reported in Appendix A.

3.2 Hidden-State Boundary

For each model and layer ℓ, we extract the residual-stream activation at the final prompt token. Let $\mu_\ell^{\mathcal{F}}$ denote the mean activation over factual prompts and $\mu_\ell^{\mathcal{U}}$ the mean activation over uncertain prompts, where uncertain is the union of hallucination and impossible prompts unless otherwise specified. The answerability boundary direction is

$$\mathbf{b}_\ell = \frac{\mu_\ell^{\mathcal{U}} - \mu_\ell^{\mathcal{F}}}{\|\mu_\ell^{\mathcal{U}} - \mu_\ell^{\mathcal{F}}\|}. \tag{5}$$

The boundary norm $\|\mu_\ell^{\mathcal{U}} - \mu_\ell^{\mathcal{F}}\|$ measures class separation. Boundary stability is the cosine similarity between adjacent-layer boundary directions. The intervention layer ℓ^* is chosen from validation activations as the first layer where the boundary is stable and topological fragmentation begins sustained growth. In the original two-model intervention setting this gives layer 16 for Llama-3.2-3B and layer 20 for Qwen-2.5-3B; the comprehensive sweeps use the corresponding model-specific layer selected by the same rule.

3.3 Prediction and Input Projectors

Let $\Pi_{\text{pred}}^{(k)}$ be the orthogonal projector onto the top k right singular vectors of W_U. This is the prediction-aligned subspace: directions in the residual stream that are most visible to the output map. Let $\Pi_{\text{input}}^{(k)}$ be the projector onto the top k principal components of residual-stream activations at the same layer. This is the input-aligned subspace: directions that carry the largest activation variance.

For a direction $\mathbf{u}$, we measure its relative subspace fraction by

$$\text{RSF}(\mathbf{u}, \Pi) = \frac{\|\Pi\mathbf{u}\|^2}{\|\mathbf{u}\|^2}. \tag{6}$$

We also report the visibility

$$\text{vis}_m(\mathbf{u}) = \frac{\|P_{\mathcal{V}_m}\mathbf{u}\|}{\|\mathbf{u}\|}, \tag{7}$$

where $\mathcal{V}_m$ is the span of the top m right singular vectors of W_U. In the main visibility table we use $m = 64$. The two measurements are related but not identical: visibility is a norm overlap with the top unembedding directions, while RSF is a squared fraction used to compare input and prediction projectors.

3.4 Detection Probes and Output Baselines

We train a logistic regression probe on hidden states at ℓ^* to classify factual prompts versus uncertain prompts. The probe has no access to logits, generated text, or prompt metadata. It tests whether answerability is linearly available in the residual stream.

We compare the probe with two output-only uncertainty baselines. The first is inverse maximum probability, $1 - \max_i p_i$, computed from the next-token distribution. The second is token entropy, $-\sum_i p_i \log p_i$. Both are evaluated as detectors for hallucination or impossible prompts versus factual prompts using AUROC. These

baselines ask whether the model's own output distribution expresses the uncertainty that the hidden-state probe detects.

3.5 Internal Geometry

We use local intrinsic dimensionality (LID) and spectral statistics to measure the geometry of class-conditional activations. For a point x with k nearest-neighbor distances $r_1 \leq \cdots \leq r_k$, the maximum-likelihood LID estimate is

$$\mathrm{LID}(x) = -\left(\frac{1}{k}\sum_{i=1}^{k}\log\frac{r_i}{r_k}\right)^{-1}. \tag{8}$$

We use $k = \min(20, N-2)$ and add small Gaussian noise to avoid degeneracy from near-duplicate activations. We also compute class-conditional covariance spectra, isotropy, spectral entropy, and the number of principal components required to explain a fixed fraction of variance.

To test whether uncertainty forms one coherent state or many disconnected regions, we compute persistent homology on activations near the answerability boundary. Points are selected by distance to the boundary, a Vietoris–Rips complex is constructed over a range of distance scales, and Betti numbers are recorded. β_0 counts connected components; large late-layer β_0 indicates fragmentation. Non-zero β_1 indicates loops and non-convex structure.

3.6 Causal Interventions

The readout bypass tests whether the internal signal can drive refusal when connected to the logits. Let $q_\ell = P(\text{uncertain} \mid \mathbf{h}^{(\ell)})$ be the logistic-probe confidence at the intervention layer. Let $\mathbf{r}_{\text{refuse}} \in \mathbb{R}^V$ be a sparse logit direction supported on refusal and uncertainty tokens. We modify the final logits by

$$\mathbf{z}_{\text{bypass}} = \mathbf{z}_{\text{model}} + \gamma q_\ell \mathbf{r}_{\text{refuse}}. \tag{9}$$

The gain γ is swept over $\{0, 1, 2, 5, 10, 20, 50\}$. We report refusal rates separately on factual, hallucination, and impossible prompts.

The boundary steering intervention tests whether the answerability boundary is a causal direction. We add a signed multiple of the boundary displacement to the residual stream:

$$\mathbf{h}^{(\ell^*)} \leftarrow \mathbf{h}^{(\ell^*)} + \alpha\left(\mu^{\mathcal{U}}_{\ell^*} - \mu^{\mathcal{F}}_{\ell^*}\right), \tag{10}$$

with $\alpha \in \{-2, -1, -0.5, 0, 0.5, 1, 2\}$, and then we measure the fraction of generations that change relative to the unedited model.

The manifold repair intervention tests whether uncertain states are a simple linear displacement away from the factual manifold. We compute a PCA subspace $S_{\mathcal{F}}$ from factual activations and softly project uncertain states toward it:

$$\mathbf{h}' = (1-\lambda)\mathbf{h} + \lambda\left[\mathrm{Proj}_{S_{\mathcal{F}}}(\mathbf{h} - \mu^{\mathcal{F}}) + \mu^{\mathcal{F}}\right], \tag{11}$$

with $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$. If uncertain states were linearly displaced factual states, this projection should improve generation. The loop rate is the rate of degenerate repeated outputs under the intervention.

4 Results

4.1 Answerability Is Detected Internally but Weakly Expressed at the Output

The central result is that hidden states carry a much better answerability signal than the output distribution. Table 1 shows that a linear probe reliably distinguishes uncertain prompts from factual prompts across model families. The corresponding output-only baselines are substantially weaker. Inverse maximum probability and entropy usually remain near the weak-detector regime. Thus the model's residual stream contains information

Table 1: Detection and routing geometry. Boundary quantities are computed at the intervention layer using the factual split and the uncertain split, where uncertain denotes hallucination plus impossible prompts. Visibility is measured at $m = 64$ using the top right singular vectors of W_U. RSF is the squared fraction of the boundary direction in the indicated projector. The AUROC columns evaluate binary detection of uncertain prompts versus factual prompts: a hidden-state linear probe, inverse maximum next-token probability, and next-token entropy.

Model	$\mathrm{vis}_{64}(\mathbf{b})$	$\mathrm{vis}_{64}(\mathbf{b}_{\mathrm{pred}})$	Random	$\mathrm{RSF}_{\mathrm{pred}}$	$\mathrm{RSF}_{\mathrm{input}}$	Input/Pred	Probe	$1 - \max p$	Entropy
Llama-3.1-8B	0.2829	0.3723	0.1426	0.1652	0.9982	6.04	0.9784	0.6828	0.6925
Llama-3.2-3B-IT	0.1563	0.4037	0.1508	0.1057	0.9997	9.46	0.9809	0.5634	0.5682
Llama-3.2-3B	0.1948	0.4178	0.1512	0.0965	0.9987	10.35	0.9384	0.7054	0.7106
Mistral-7B-v0.1	0.7738	0.5782	0.1242	0.6698	1.0000	1.49	0.9868	0.6928	0.6922
Qwen-2.5-7B	0.5510	0.6979	0.1227	0.3272	1.0000	3.06	0.9244	0.5483	0.5842
Qwen-2.5-3B-IT	0.2384	0.3359	0.1646	0.1519	0.9999	6.58	0.9334	0.6100	0.6146
Qwen3-8B	0.0857	0.3088	0.1219	0.0113	1.0000	88.73	0.9781	0.6066	0.6505
Qwen3-32B	0.1719	0.6610	0.1058	0.0627	1.0000	15.94	0.8600	0.5731	0.6121

about whether the prompt is answerable, but the next-token distribution is a poor public readout of that information.

The same table shows the subspace structure of the boundary. The boundary is almost fully contained in the input projector for every model tested: $\mathrm{RSF}(\mathbf{b}, \Pi_{\mathrm{input}})$ is approximately one. The prediction-side fraction is smaller and varies substantially across architectures. In Llama and Qwen instruction-tuned models, the input-to-prediction ratio is typically between about 6 and 16; in Qwen3-8B it is nearly 89. Mistral is the main exception: its boundary has a larger prediction-side component, and the ratio is correspondingly smaller. This exception shows that not that all uncertainty directions are literally absent from the prediction subspace. Rather, that answerability is primarily organized as input geometry, while output expression depends on how much of that geometry is converted into prediction geometry.

4.2 The Boundary Grows Internally Without Becoming a Stable Output Variable

The residual-stream boundary is not a small or fragile effect. The boundary norm increases with depth and is stable across adjacent layers. Table 2 summarizes the late-layer geometry. Boundary norms vary by architecture, but the qualitative pattern is consistent: the representation separates factual and uncertain prompts, local Fisher sensitivity is modest relative to the size of the boundary, Hessian curvature along the boundary is near zero, and Jacobian amplification is above one. In plain terms, the network can amplify the answerability direction internally while leaving the local output landscape almost flat along that direction.

Intrinsic-dimensionality and spectral measurements give the same picture. Uncertain prompts occupy higher-dimensional regions than factual prompts, and their covariance spectra are less concentrated. Persistent homology adds a second fact: the uncertain region does not collapse into one abstention state. β_0 grows through depth, indicating that the uncertainty manifold fragments into many connected components. This matters for interventions. If uncertainty were a single linear displacement away from factual processing, a projection back onto the factual subspace should help. The repair experiments below show that it does not.

4.3 Component-Level Mechanisms

The same routing pattern appears when the residual update is decomposed into attention and MLP contributions. Late-layer movement along the hallucination boundary is dominated by MLP outputs rather than attention outputs. In Qwen-2.5, for example, hallucinatory inputs at layer 27 show substantial alignment with MLP updates but little alignment with attention updates. This suggests that, once contextual grounding is weak, associative transformations in the MLP can move the residual stream along the uncertainty direction without a corresponding attention-mediated correction.

Attention statistics support this interpretation. In Mistral-7B-v0.1, factual inputs at a late layer concentrate attention strongly on the sink position, whereas hallucinatory inputs distribute attention more diffusely. The

Table 2: Late-layer geometric summary on the impossible boundary. Boundary stability is high in all models and omitted for compactness.

Model	Norm	Fisher	Hessian	Jac. Amp.
Llama-3.1-8B	7.52	0.076	0.020	1.237
Llama-3.2-3B	6.44	0.073	0.021	1.211
Llama-3.2-3B-IT	6.59	0.547	0.011	1.279
Mistral-7B-v0.1	23.02	1.239	-0.010	1.122
Qwen-2.5-7B	33.59	0.325	0.000	1.606
Qwen-2.5-3B-IT	19.72	4.421	-0.004	1.566
Qwen3-8B	69.35	0.762	0.003	1.996
Qwen3-32B	97.86	1.451	-0.002	2.071

model is not simply inactive on hallucinations; rather, attention remains active but less anchored. Final-layer activation statistics point to two output regimes. Some hallucinations correspond to high-entropy guessing, while others are confident commitments driven by sparse outlier features. In both cases, the component-level evidence is consistent with the anisotropic account: the model has internal structure associated with uncertainty, but the downstream route from that structure to calibrated expression is unreliable. Details and auxiliary tables are given in Appendix B.

4.4 Training Dynamics

Developmental checkpoints show that the answerability geometry is not an artifact of a final trained model. In OLMo and Pythia checkpoints, factual and hallucinatory prompts are relatively similar early in training, when residual-stream representations are still weakly organized. As training proceeds, factual representations compress and uncertain representations separate into higher-dimensional regions. By later checkpoints, impossible and hallucination prompts occupy distinct regions with larger intrinsic dimension than factual prompts. This trajectory is consistent with the anisotropic routing account: pretraining organizes the residual stream into structured input and prediction geometries, but it does not necessarily create a stable output target for abstention.

The checkpoint evidence should be interpreted as developmental support rather than a complete training theory. The available public checkpoint suites do not allow every intervention to be repeated at every scale, and instruction tuning can change local sensitivity in model-dependent ways. What we can say is that the detection-expression gap emerges during the *formation* of residual-stream geometry, not only after alignment or prompting. Additional checkpoint details are provided in Appendix B.

4.5 Instruction Tuning Changes Sensitivity but Not the Basic Routing Problem

Instruction-tuned models show larger local Fisher sensitivity in some cases, especially Llama-3.2-3B-IT and Qwen-2.5-3B-IT. This is expected: instruction tuning contains demonstrations of refusal and hedging, so it can increase the local coupling between answerability and output. But the effect is incomplete. The hidden-state probe remains much stronger than output entropy or maximum probability, and the boundary remains dominated by input geometry. Instruction tuning therefore appears to add partial output sensitivity without fully changing the routing structure. So it is not the case that aligned models never learn to refuse, more that ordinary output behavior still underuses a robust answerability signal that is already available internally.

5 Causal Experiments

The causal experiments ask whether the boundary can control behavior, whether a direct readout can express the hidden signal, and whether a simple geometric repair can recover coherent text generation.

Table 3: Readout bypass refusal rates. Factual refusal at $\gamma = 50$ remains zero for most models and small for Qwen3-32B. Hallucination and impossible columns show baseline, moderate-gain, and high-gain refusal.

Model	Factual $\gamma = 50$	Hall. $\gamma = 0$	Hall. $\gamma = 20$	Hall. $\gamma = 50$	Imp. $\gamma = 0$	Imp. $\gamma = 20$	Imp. $\gamma = 50$
Llama-3.1-8B	0.000	0.000	0.492	0.625	0.008	0.992	1.000
Llama-3.2-3B	0.000	0.000	0.417	0.542	0.000	0.992	1.000
Llama-3.2-3B-IT	0.000	0.125	0.267	0.675	0.058	0.467	1.000
Mistral-7B-v0.1	0.000	0.008	0.450	0.650	0.000	0.975	1.000
Qwen-2.5-7B	0.000	0.008	0.658	0.725	0.000	1.000	1.000
Qwen-2.5-3B-IT	0.000	0.450	0.450	0.808	0.125	0.125	0.992
Qwen3-8B	0.000	0.008	0.308	0.633	0.008	0.725	1.000
Qwen3-32B	0.050	0.025	0.250	0.692	0.008	0.650	0.983

Table 4: Boundary steering output-change rate on hallucination prompts. The $\alpha = 0$ column is omitted because it is zero by definition.

Model	$\alpha = -2$	$\alpha = -1$	$\alpha = -0.5$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
Llama-3.1-8B	0.767	0.508	0.325	0.233	0.433	0.658
Llama-3.2-3B	0.842	0.558	0.383	0.333	0.583	0.858
Llama-3.2-3B-IT	0.808	0.525	0.350	0.358	0.542	0.725
Mistral-7B-v0.1	0.983	0.783	0.475	0.650	0.817	0.992
Qwen-2.5-7B	1.000	1.000	0.983	0.975	1.000	1.000
Qwen-2.5-3B-IT	0.742	0.517	0.367	0.308	0.483	0.717
Qwen3-8B	1.000	0.950	0.658	0.642	0.808	1.000
Qwen3-32B	1.000	1.000	1.000	1.000	1.000	1.000

5.1 Readout Bypass

The readout bypass converts hidden-state probe confidence into extra logit mass on refusal tokens. Table 3 shows three effects. First, small gains have little effect; the model does not spontaneously use the signal unless the bypass has enough strength. Second, high gain produces strong and mostly selective refusal on uncertain prompts, especially impossible prompts. At $\gamma = 50$, impossible-prompt refusal reaches approximately one in all models. Third, factual refusal remains zero in most models and small in Qwen3-32B even at the largest gain. This is the cleanest causal evidence for the detection-expression gap: the signal is present and usable, but it requires an explicit output route.

Hallucination prompts show a softer response than impossible prompts. This is also expected. Impossible prompts are unanswerable as stated, while hallucination prompts are often answerable in principle but difficult for the model. The probe confidence therefore separates them less sharply, and the bypass produces intermediate refusal rates rather than universal abstention.

5.2 Boundary Steering

Boundary steering asks whether the answerability direction is merely diagnostic or actually causal. The answer is causal. Table 4 reports output-change rates on hallucination prompts under signed boundary injections. The rates grow with $|\alpha|$ and are high at $\alpha = \pm 2$ for all models. Qwen-2.5-7B, Qwen3-8B, and Qwen3-32B are especially sensitive, reaching near-complete output change for moderate or large steering. Llama and instruction-tuned Qwen are less sensitive but still show strong effects.

The sign asymmetry is notable. Negative and positive steering both change outputs, but not always equally. This indicates that the boundary does not act as a calibrated refusal dial by itself. It is a direction that moves the representation into a different processing regime. Whether that regime produces hedging, a different confident answer, or degeneration depends on how the downstream model maps the perturbed state into prediction space.

Table 5: Manifold repair loop rate under projection toward the factual PCA subspace. Projection does not reduce degeneration and often increases it at high λ.

Model	Split	$\lambda = 0$	$\lambda = 0.25$	$\lambda = 0.50$	$\lambda = 0.75$	$\lambda = 1.00$
Llama-3.1-8B	Hallucination	0.000	0.000	0.008	0.100	0.008
	Impossible	0.000	0.000	0.000	0.058	0.000
Llama-3.2-3B	Hallucination	0.000	0.000	0.008	0.233	0.067
	Impossible	0.000	0.000	0.025	0.125	0.042
Llama-3.2-3B-IT	Hallucination	0.000	0.000	0.000	0.242	0.033
	Impossible	0.000	0.000	0.000	0.383	0.042
Mistral-7B-v0.1	Hallucination	0.000	0.000	0.042	0.400	0.000
	Impossible	0.000	0.000	0.008	0.408	0.000
Qwen-2.5-7B	Hallucination	0.000	0.000	0.000	0.442	0.000
	Impossible	0.017	0.033	0.008	0.467	0.142
Qwen-2.5-3B-IT	Hallucination	0.000	0.000	0.000	0.042	0.658
	Impossible	0.000	0.000	0.000	0.033	0.775
Qwen3-8B	Hallucination	0.000	0.000	0.000	0.950	1.000
	Impossible	0.000	0.000	0.000	0.975	1.000
Qwen3-32B	Hallucination	0.000	0.000	0.000	0.333	0.000
	Impossible	0.000	0.000	0.000	0.275	0.000

5.3 Manifold Repair

The repair experiment provides an informative negative result. If uncertain representations were simply factual representations plus a linear displacement, projection onto the factual PCA subspace should reduce degeneration. It does not. Table 5 shows that low projection strengths usually leave loop rate unchanged, while stronger projection often makes generation worse. The effect is striking in Qwen3-8B, where $\lambda = 0.75$ and $\lambda = 1.0$ produce loop rates close to one. Qwen3-32B shows a milder but similar degradation at $\lambda = 0.75$, and Qwen-2.5-3B-IT shows the same failure mode at full projection.

This supports the topological interpretation. The uncertain region is not a convex cloud that can be moved back to the factual manifold by a linear map. It is fragmented. Projection discards information needed for coherent generation without reconstructing a valid factual state. The result also clarifies the role of the readout bypass: successful intervention requires an output route for uncertainty, not a repair that pretends uncertainty is corrupted factuality.

6 Discussion

The results suggest that hallucination is not only a failure of factual storage, it's also a failure of routing. The model builds a representation that separates answerable from uncertain prompts, but this representation is organized primarily as input geometry. Output behavior depends on prediction geometry. When the training objective does not require a stable abstention target, the transformation from input-side answerability to prediction-side refusal remains incomplete.

This interpretation also explains why different interventions behave differently. A probe works because the signal is linearly present. A bypass works because it supplies the missing output route. Steering works because the boundary is a real causal axis, but it does not necessarily produce clean refusal because the downstream mapping is not calibrated for that axis. Repair fails because uncertainty is not a simple linear error in factual representation space. The experiments are therefore mutually constraining: each intervention succeeds or fails in the way predicted by the anisotropic routing account.

The architectural variations are important too: Mistral has relatively high raw visibility of the boundary in the top unembedding directions, while Qwen3 models show much weaker prediction-side RSF, especially Qwen3-8B. Yet these models all show a gap between hidden-state detection and output uncertainty metrics. This means that no single scalar visibility measurement is sufficient. The phenomenon is better described

as a mismatch between the geometry that detects answerability and the geometry that expresses token commitments.

7 Related Work

Hallucination and factuality failures have been studied in summarization, question answering, and open-ended generation (Kalai et al., 2025; Lin et al., 2022; Ji et al., 2023; Zhang et al., 2025). Many mitigation methods treat hallucination as a knowledge-access problem. Retrieval augmentation supplies external documents at inference time (Lewis et al., 2020); knowledge editing and fine-tuning modify model parameters or behavior (Kandpal et al., 2023; Lv et al., 2024); decoding and verification methods attempt to reduce unsupported generations (Lee et al., 2022). Our results are compatible with these approaches but address a different failure mode: even when the model internally represents that a prompt is unreliable, the representation may not control output behavior.

Several recent works show that hidden states contain truthfulness or uncertainty information not directly visible in the generated answer. Semantic entropy detects confabulation by measuring uncertainty over meanings rather than tokens (Farquhar et al., 2024). Other work trains probes or unsupervised detectors on hidden representations to predict truthfulness or hallucination (Azaria & Mitchell, 2023; Chen et al., 2024; Su et al., 2024; Sriramanan et al., 2024). Neuron-level and sparse-feature analyses identify localized components associated with hallucination or over-compliance (Gao et al., 2025; Suresh et al., 2026). Our contribution is to connect these detection results to output geometry: the signal is not merely present, it is routed through a subspace different from ordinary prediction.

The paper also relates to work on internal knowledge and output mismatch. Orgad et al. (2025) show that models may encode correct answers internally while generating incorrect ones. Cohen et al. (2024) propose explicit uncertainty expression through an "IDK" token, which is close in spirit to our readout bypass but implemented as a training intervention. Ren et al. (2024) analyze fine-tuning dynamics that strengthen hallucination through inappropriate transfer across contexts. Our anisotropic account gives a geometric mechanism for such mismatches: internal features can be linearly available without being aligned to the output-sensitive directions used for token commitment.

Finally, our analysis builds on residual-stream Basile et al. (2024) and subspace views of transformer computation. The distinction between input-aligned and prediction-aligned geometry follows the anisotropic gradient-accumulation perspective, in which activation covariance shapes read-side structure while sharp output gradients shape write-side prediction geometry. We apply that framework to answerability and hallucination rather than to alignment deltas or token suppression.

8 Limitations

The experiments show a robust detection-expression gap, but they do not prove that this mechanism explains every hallucination. Some hallucinations are genuine knowledge failures, for which, retrieval or additional training can still be the right remedy. Our results are narrower: there exists a class of failures in which answerability is represented internally but not routed to output behavior.

The causal interventions are diagnostic rather than deployable solutions. The readout bypass uses a hand-specified refusal-token direction and a gain parameter. It demonstrates that the hidden signal can drive abstention, but it is not a complete calibration method. Boundary steering is also not a clean control interface: it changes outputs reliably, but the qualitative effect depends on the model. Manifold repair uses a linear factual subspace and therefore only tests one simple repair hypothesis.

The evaluation is limited to the model families and prompt distributions studied here. The prompt splits are manually verified, but refusal and looping metrics depend on operational classifiers. Future work should test multilingual prompts, tool-use settings, retrieval-augmented systems, and training interventions that explicitly rotate answerability features into prediction-aligned directions. A stronger test would train models with controlled abstention targets and measure whether $\mathrm{RSF}(\mathbf{b}, \Pi_{\mathrm{pred}})$ increases while hallucination decreases.

9 Conclusion

We propose an anisotropic routing account of hallucination. Transformer residual streams can encode answerability as a strong input-aligned feature, while the output distribution remains governed by prediction-aligned directions shaped by next-token training. This creates a detection-expression gap: the model can know, in its hidden states, that a prompt is unreliable without expressing that uncertainty in its answer.

The experiments support this account. Hidden-state probes detect uncertain prompts much better than output confidence or entropy. The answerability boundary lies almost entirely in the input manifold and has weaker, architecture-dependent, prediction alignment. A readout bypass induces selective refusal, boundary steering changes outputs, and linear manifold repair fails. These results suggest that reducing hallucination requires more than adding knowledge. It requires training or architectural mechanisms that route internal answerability signals into the output geometry where decisions are expressed.

References

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Franz Aurenhammer, Evanthia Papadopoulou, and Martin Suderland. Piecewise-linear farthest-site voronoi diagrams. In *32nd International Symposium on Algorithms and Computation (ISAAC 2021)*, pp. 30–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2021.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Lorenzo Basile, Valentino Maiorca, Luca Bortolussi, Emanuele Rodolà, and Francesco Locatello. Residual transformer alignment with spectral decomposition. *arXiv preprint arXiv:2411.00246*, 2024.

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. *arXiv preprint arXiv:2402.03744*, 2024.

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.

Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. I don't know: Explicit modeling of uncertainty with an [idk] token. *Advances in Neural Information Processing Systems*, 37:10935–10958, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *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. 55–65, 2019.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

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.

William H Guss and Ruslan Salakhutdinov. On characterizing the capacity of neural networks using algebraic topology. *arXiv preprint arXiv:1802.04443*, 2018.

Christoph Hofer, Roland Kwitt, Marc Niethammer, and Andreas Uhl. Deep learning with topological signatures. *Advances in neural information processing systems*, 30, 2017.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning*, pp. 15696–15707. PMLR, 2023.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation. *Advances in neural information processing systems*, 35:34586–34599, 2022.

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. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Kai Lv, Yuqing Yang, Tengxiao Liu, Qipeng Guo, and Xipeng Qiu. Full parameter fine-tuning for large language models with limited resources. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8187–8198, 2024.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

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 *International Conference on Learning Representations*, volume 2025, pp. 66880–66913, 2025.

Stefano Recanatesi, Matthew Farrell, Madhu Advani, Timothy Moore, Guillaume Lajoie, and Eric Shea-Brown. Dimensionality compression and expansion in deep neural networks. *arXiv preprint arXiv:1906.00443*, 2019.

Richard Ren, Steven Basart, Adam Khoja, Alexander Pan, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Gabriel Mukobi, Ryan H Kim, et al. Safetywashing: Do ai safety benchmarks actually measure safety progress? *Advances in Neural Information Processing Systems*, 37:68559–68594, 2024.

Bastian Rieck, Matteo Togninalli, Christian Bock, Michael Moor, Max Horn, Thomas Gumbsch, and Karsten Borgwardt. Neural persistence: A complexity measure for deep neural networks using algebraic topology. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=ByxkijC5FQ`.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216, 2024.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14379–14391, 2024.

Praneet Suresh, Jack Stanley, Sonia Joseph, Luca Scimeca, and Danilo Bzdok. From noise to narrative: Tracing the origins of hallucinations in transformers. *Advances in Neural Information Processing Systems*, 38:22465–22502, 2026.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36:51234–51252, 2023.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418, 2025.

A Dataset

A.1 Dataset Controls

A possible confound is that the factual and uncertain splits differ in surface complexity rather than answerability. Table 6 reports syntactic and readability statistics for all three evaluation sets. The factual and impossible splits are closely matched across all measured variables. In fact, the impossible split is slightly simpler than the factual split on dependency depth, Flesch–Kincaid grade level, and reading ease. This is the opposite direction predicted by an explanation in which impossible prompts induce different geometry because they are syntactically more complex.

The hallucination split is longer and has slightly more clauses, reflecting the use of long-tail factual questions. This difference does not explain the main factual-versus-impossible boundary, which is the cleanest test of answerability while controlling for surface form. Hidden-state probes are still trained on representations that may contain syntactic information, so these controls do not prove that syntax plays no role. They do show that the measured geometric separation is not reducible to the standard surface-complexity variables in Table 6. The causal interventions further weaken a surface-only account: directly routing the hidden-state signal to refusal logits selectively changes behavior on uncertain prompts, while projection onto the factual subspace does not repair generation.

Table 6: Syntactic and readability controls for the three evaluation sets. Values are mean ± standard deviation.

Metric	Factual	Impossible	Hallucination
Token count	6.34 ± 1.86	6.45 ± 2.36	8.58 ± 4.21
Dependency tree depth	2.56 ± 0.62	2.41 ± 0.57	2.77 ± 0.46
Clause count	1.09 ± 0.29	1.08 ± 0.28	1.33 ± 0.62
Flesch–Kincaid grade	4.29 ± 3.72	4.08 ± 3.64	4.75 ± 3.33
Flesch–Kincaid reading ease	75.53 ± 27.24	77.16 ± 25.09	75.23 ± 22.69

A.2 Datasets samples

Here we present a sample of the datasets we generated using Claude and Gemini, and that we also manually checked. In the supplementary materials you can find the full datasets.

The evaluation datasets were constructed to isolate distinct epistemic regimes rather than to maximize difficulty per se. The Factual Knowledge dataset, as we can see in Table 7 was intentionally composed of simple, high-confidence questions (e.g., capitals, authorship, basic scientific facts) on which contemporary models achieve consistently high accuracy. This ensures that representations in this regime reflect genuine model knowledge, providing a stable geometric baseline for comparison.

In contrast, the Impossible Questions dataset, as we can see in Table 8 was explicitly designed to elicit confabulation by presenting prompts that cannot be answered meaningfully. These include both false-presupposition questions (e.g., anachronistic or logically inconsistent premises) and context-deficient queries that lack sufficient information for resolution. Multiple prompt styles were used to avoid stylistic bias and to probe whether models recognize unanswerability across diverse linguistic forms.

For diffusion models, the Paradoxical Prompts dataset targets well-documented failure modes by requesting internally inconsistent or combinatorially constrained scenes, such as precise object counts, specific color assignments, legible text, or coherent spatial layouts. These prompts describe plausible scenarios but impose requirements that current text-to-image systems systematically fail to satisfy, allowing us to probe uncertainty-related geometry in continuous generative settings.

Together, these datasets enable controlled comparisons between grounded knowledge, meaningful but unanswerable inputs, and hallucination-prone regimes, ensuring that observed geometric effects reflect epistemic structure rather than generic task difficulty.

Question	Answer	Category
What is the capital of Greece?	Athens	Capitals
What is the capital of Poland?	Warsaw	Capitals
What is the capital of Sweden?	Stockholm	Capitals
Who wrote *Wuthering Heights*?	Emily Brontë	Authors
Who wrote *Frankenstein*?	Mary Shelley	Authors
Who wrote *Dracula*?	Bram Stoker	Authors
Who founded Twitter?	Jack Dorsey	Companies
Who founded Nike?	Phil Knight	Companies
Who founded Walmart?	Sam Walton	Companies
What is the unit of force?	Newton	Science
What is the unit of energy?	Joule	Science

Table 7: Representative examples from the Factual Knowledge dataset. Each entry consists of a question, its verified answer, and semantic category.

Impossible questions Dataset Samples
Wrong premise questions
What is the ISBN of Albert Einstein's 1957 cookbook 'Relativity Recipes'?
What social media platform did Julius Caesar use to announce his victories?
What was George Washington's TikTok username?
What is the patent number for Leonardo da Vinci's microwave oven?
How tall was Napoleon's pet dinosaur?
What is the entrance fee for the Forbidden City in Rio de Janeiro?
How fast can a goblin run in kilometers per hour?
What was Genghis Khan's email address?
No context questions
Who won the Super Bowl that year?
How much does it cost?
Did the database sync?
Is the API public?
Does it support wireless charging?
Did the patch fix the bug?
How many megapixels is the camera?
Is the server online?
Is the warranty transferable?
What is the discount code?

Table 8: Representative examples from the Impossible questions dataset.

B Component and Developmental Analyses

B.1 Component-Level Measurements

For each MLP neuron, we compute a class-selectivity score

$$s_i = \frac{\mu_i^{\mathcal{U}} - \mu_i^{\mathcal{F}}}{\sigma_i^{\mathcal{U}} + \sigma_i^{\mathcal{F}}}, \tag{12}$$

where the means and variances are estimated with Welford's online algorithm for memory efficiency (**?**). For each attention head, we compute the entropy of its attention distribution by input class and measure class

Paradox Prompts Dataset Samples
A digital clock display showing exactly 09:07 (leading zero required).
"A photo of exactly 7 apples in a straight line, all visible and not overlapping."
A desk with exactly 9 paperclips arranged in a 3×3 grid.
A beach scene with exactly 5 umbrellas, each a different color, all fully visible.
A red cube, a blue sphere, and a green pyramid, left to right in that order.
"A dog wearing sunglasses and a cat wearing a bowtie, plus another dog wearing a bowtie and another cat wearing sunglasses (four animals total)."
A person holding a sign above their head; the sign is readable and not tilted.
A floating book that is visibly not supported by strings, hands, or surfaces (clean background).

Table 9: Representative examples from the Paradox prompts dataset for image generation.

divergence in attention entropy and sink mass. These measurements are not used to define the boundary or train the probe. They provide a component-level account of where the boundary is amplified and where grounding signals weaken.

The strongest qualitative pattern is an asymmetry between attention and MLP contributions. In Qwen-2.5, hallucinatory inputs at layer 27 show strong alignment with MLP outputs, with mlp_align ≈ 0.268, but minimal alignment with attention outputs, with attn_align ≈ 0.017. This suggests that late movement along the hallucination direction is driven primarily by associative MLP transformations. Attention, by contrast, provides little corrective movement along the boundary at that layer.

Attention sink behavior provides a complementary signal. In Mistral-7B-v0.1 at layer 31, factual inputs allocate approximately 88% of attention mass to the sink position, while hallucinatory inputs allocate approximately 49%. Hallucinations therefore do not correspond to absent attention. They correspond to diffuse attention: attention remains active, but is less anchored to a stable grounding pattern. Final-layer activation statistics further suggest two regimes. Some hallucinations produce high-entropy guessing; others produce confident but incorrect commitments when fragmented representations align with strong priors. Both regimes can exhibit extreme activation kurtosis, with correct predictions typically in the range 7–16 and hallucinated outputs reaching approximately 280–350. This is consistent with hallucinated commitments being driven by sparse outlier features rather than distributed consensus.

B.2 Developmental Checkpoints

We apply the same metric suite to available OLMo and Pythia checkpoints to track when the answerability geometry emerges. Early in training, factual and hallucinatory inputs are geometrically close: at OLMo step 1,000, the factual LID is approximately 12.6 and the hallucination LID is approximately 14.9. Representations at this stage are weakly organized, and hallucinations resemble unstructured noise more than a stable failure mode.

As training progresses, factual representations compress and uncertainty-related representations become more separated. By step 10,000, factual LID drops to approximately 10.5, indicating that answerable prompts are beginning to occupy a more compact region of the residual stream. Later, the model increasingly separates known from unknown. By step 100,000, factual LID is approximately 14.8, hallucination LID is approximately 22.0, and impossible-prompt LID is approximately 25.0. Hallucination representations retain partial structure, but they remain too diffuse to support reliable output commitment. These trajectories support the view that the detection-expression gap forms during pretraining as residual-stream geometry becomes anisotropic.

Table 10: Intrinsic dimension, isotropy, and spectral entropy aggregated by model and input class.

Model	Input	LID	Isotropy	Entropy
Llama-3.1-8B	Factual	11.79	0.659	319.9
	Impossible	18.05	0.519	664.0
	Hallucination	15.09	0.697	618.8
Llama-3.2-3B	Factual	11.70	0.581	312.9
	Impossible	17.83	0.545	633.0
	Hallucination	14.53	0.704	578.6
Llama-3.2-3B-IT	Factual	11.97	0.609	314.7
	Impossible	17.59	0.618	635.4
	Hallucination	14.56	0.605	596.8
Qwen-2.5-7B	Factual	9.71	0.582	265.7
	Impossible	14.27	0.406	554.4
	Hallucination	12.62	0.511	507.1
Qwen-2.5-3B-IT	Factual	9.83	0.581	250.4
	Impossible	14.16	0.462	507.9
	Hallucination	12.48	0.548	470.9
Qwen3-32B	Factual	10.41	0.541	279.2
	Impossible	16.76	0.524	636.5
	Hallucination	14.70	0.543	579.5
Qwen3-8B	Factual	11.21	0.548	282.7
	Impossible	15.63	0.617	603.1
	Hallucination	13.96	0.562	560.1

C Additional Geometry Tables

D Implications for Intervention Design

The experiments suggest that mitigation should focus on coupling internal answerability detection to output behavior. The following intervention classes are natural consequences of the measurements, but we treat them as hypotheses for future work rather than evaluated methods.

At pretraining time, one could add objectives that make abstention a stable target when inputs are underspecified or false-premised. The geometric version of this idea is to encourage uncertain representations to collapse toward a compact uncertainty schema rather than fragmenting across disconnected components. A dedicated uncertainty token could also be explicitly anchored to residual-stream directions that show high answerability signal but weak prediction alignment.

At alignment time, preference optimization could be made geometry-aware. Instead of treating refusals uniformly, updates could be weighted by boundary visibility, Fisher sensitivity, or probe confidence, targeting examples where detection is present but output coupling is weak. Component-level measurements suggest another possibility: late-layer MLP contributions could be regularized when attention is diffuse and activation kurtosis is extreme, reducing the tendency for sparse associative features to dominate output commitment.

At inference time, the readout bypass provides the simplest proof of concept: a probe can route hidden-state answerability to refusal logits. Less direct variants could monitor boundary projection, entropy, and activation kurtosis jointly, then adapt decoding or trigger clarification. These approaches should be calibrated carefully, since boundary steering shows that the answerability direction is causal but not by itself a clean refusal dial.

Table 11: Attention and logit statistics by model and input class. Output confidence remains similar across input classes even when internal geometry differs.

Model	Input	Attn. Ent.	Sink Mass	Res. Norm	Confidence
Llama-3.1-8B	Factual	0.96	0.716	18.5	0.047
	Hallucination	1.06	0.701	18.8	0.024
	Impossible	1.01	0.703	19.0	0.022
Llama-3.2-3B	Factual	0.87	0.736	16.7	0.048
	Hallucination	1.01	0.709	17.5	0.020
	Impossible	0.96	0.713	17.6	0.025
Llama-3.2-3B-IT	Factual	0.85	0.756	16.1	0.057
	Hallucination	0.97	0.733	16.6	0.030
	Impossible	0.96	0.723	16.7	0.032
Mistral-7B-v0.1	Factual	1.16	0.227	50.6	0.094
	Hallucination	1.12	0.154	43.8	0.064
	Impossible	1.01	0.153	46.5	0.071
Qwen-2.5-7B	Factual	1.17	0.535	108.0	0.251
	Hallucination	1.27	0.521	106.9	0.244
	Impossible	1.17	0.540	109.1	0.268
Qwen-2.5-3B-IT	Factual	1.20	0.452	85.5	0.187
	Hallucination	1.28	0.445	86.1	0.189
	Impossible	1.21	0.458	85.4	0.212
Qwen3-32B	Factual	0.66	0.005	329.2	0.269
	Hallucination	0.72	0.005	324.4	0.266
	Impossible	0.69	0.006	341.3	0.270
Qwen3-8B	Factual	1.05	0.557	222.8	0.330
	Hallucination	1.13	0.546	224.3	0.312
	Impossible	1.09	0.549	230.3	0.324

E Image-Generation Experiments

E.1 Motivation

The main experiments study autoregressive language models, where the output map is an unembedding matrix and hallucination appears as an unsupported linguistic commitment. We also test whether the same representational pattern appears in a different generative setting: text-conditioned image generation. This test is useful because image-generation failures do not arise from next-token decoding, and the model has no natural "I do not know" output channel. If paradoxical or underspecified prompts nevertheless produce a distinctive internal geometry, this would suggest that the detection-expression gap is not merely a peculiarity of language-model vocabularies. Instead, it may reflect a broader property of generative models trained to produce outputs even when the conditioning signal is inconsistent or underspecified.

E.2 Methodology

We evaluate PixArt-Σ on a set of paradoxical image prompts and a matched control set. The paradoxical prompts target common image-generation failure modes: clocks, rendered text, object counting, spatial relations, and physically inconsistent scenes. The control prompts use similar surface form and visual domains but describe feasible scenes. Thus the comparison is designed to separate prompt content from prompt feasibility, in the same spirit as the factual-versus-impossible contrast used for language models.

For each prompt, we collect hidden activations from the denoising transformer at fixed denoising timesteps and layers. Since diffusion models do not have a final-token residual stream, we aggregate activations over

spatial tokens to obtain a prompt-level representation at each layer and timestep. We then compute the same class-conditional geometric measurements used in the language experiments: centroid separation between control and paradoxical prompts, boundary stability across depth, local intrinsic dimensionality, covariance spectra, and persistent-homology summaries near the boundary.

To characterize the temporal dynamics of these failures in a continuous output space, we also use an intermediate latent decoding protocol. During the standard 20-step DDIM denoising schedule for PixArt-Σ (XL-2-1024-MS), we intercept the latent state z_t at steps $t \in \{0, 5, 10, 15, 19\}$. At each intercept, we decode the current noisy latent through the VAE. These decoded intermediates do not represent final samples, but they provide a visual trace of how the model's denoising trajectory organizes the prompt over time. This lets us compare the geometric measurements with the visible emergence of semantic drift, compositional instability, or artifact formation.

There are two important differences from the language-model setting. First, the diffusion model's output head predicts image latents or noise residuals rather than vocabulary logits, so there is no direct analogue of refusal-token probability. Second, because the model is trained to generate an image for every prompt, the relevant behavioral failure is not overconfident text but visual artifact formation: incorrect counts, incoherent text, inconsistent spatial layouts, or physically impossible object structure. The image-generation experiment therefore tests internal detection and representational fragmentation, but not linguistic abstention.

E.3 Results

The image-generation results show the same qualitative structure observed in language models. Paradoxical prompts separate from matched controls in hidden representation space, and the separation is stable across intermediate layers. The effect is strongest for prompt classes that require precise symbolic or relational structure, such as object counting, text rendering, clock faces, and spatial relations. These are also the cases where the generated image is most likely to contain visible artifacts.

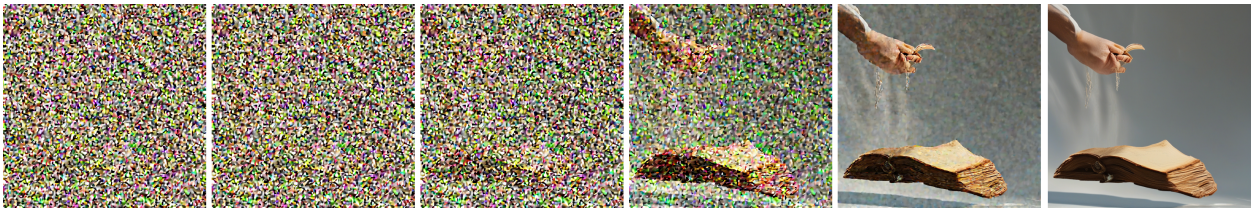
The local geometry of paradoxical prompts is higher-dimensional and less concentrated than that of matched controls. Rather than forming a single coherent "failure" state, paradoxical prompts occupy fragmented regions of activation space. This mirrors the language-model result: uncertain or unreliable inputs are internally distinguishable, but the representation does not collapse into a compact output policy. In the image model, the absence of an abstention channel means that this internal structure is expressed as visual inconsistency rather than refusal.

Persistent-homology measurements support the same interpretation. Representations near the paradoxical-control boundary develop multiple connected components as processing proceeds, indicating that different visual failure modes occupy distinct regions rather than a single unified manifold. This is consistent with the qualitative diversity of image-generation failures: a prompt involving impossible object counts, a prompt asking for coherent text, and a prompt specifying contradictory spatial relations can all be problematic, but they fail through different generative routes. Figure 2 shows representative denoising trajectories for prompts that induce these failure modes.

The intermediate decodings reveal two recurring trajectory-level patterns. In the first, the model resolves the conflict by drifting toward a high-prior visual template. For example, in Figure 2c, the denoising trajectory initially forms a conventional scene structure and only later commits to details that satisfy parts of the prompt while introducing historically or semantically implausible content. This resembles confident hallucination in language models: the model produces a coherent output, but the coherence is achieved by overriding the problematic constraint.

In the second pattern, the model fails to stabilize the global composition. Spatial or relational prompts can remain amorphous or structurally unstable through intermediate denoising steps, as illustrated in Figure 2e. This behavior is consistent with the high-dimensional and fragmented geometry measured for paradoxical prompts: the denoising trajectory does not quickly settle into a compact low-dimensional basin corresponding to a single feasible scene. The final image is therefore forced to resolve incompatible constraints through blur, object duplication, incorrect relations, or other artifacts.

(a) Intermediate steps for the prompt: “A classroom poster that reads exactly: THE QUICK BROWN FOX JUMPS OVER THE LAZY DOG.”



(b) Intermediate steps for the prompt: “A floating book that is visibly not supported by strings, hands, or surfaces (clean background).”

(c) Intermediate steps for the prompt: “A daguerreotype of Ancient Greek philosophers.”

(d) Intermediate steps for the prompt: “A courtroom scene where everyone is wearing scuba gear.”

(e) Intermediate steps for the prompt: “A cup to the left of a plate, and a bowl to the right of the plate (three objects aligned).”

Figure 2: Representative intermediate denoising trajectories for prompts that induce image-generation failures. The examples span rendered text, physical support, historical/style consistency, unusual scene composition, and spatial relations.

E.4 Conclusion

The image-generation experiment provides auxiliary evidence for the anisotropic routing account. It shows that generative models can internally distinguish feasible from problematic conditioning inputs even when their output interface does not provide a natural way to express uncertainty. In language models, this mismatch appears as confident hallucination rather than abstention. In text-to-image generation, it appears as artifact formation under paradoxical or underspecified prompts. The shared pattern is that detection-like geometry exists internally, while the training objective continues to route the model toward producing an output. This supports the broader claim that hallucination-like behavior is not only a knowledge problem, but a failure to connect internal reliability signals to an appropriate output mode.

F Extended Related Work

Intrinsic Dimensionality and Representation Geometry. A growing body of work has studied the effective dimensionality of neural representations as a lens into model behavior and generalization. Early investigations by Ansuini et al. (Ansuini et al., 2019) showed that deep networks progressively compress representations into low-dimensional manifolds, with task-relevant information concentrating in a small number of directions. Subsequent studies extended this perspective across architectures and tasks, linking intrinsic dimensionality to expressivity, robustness, and generalization (Valeriani et al., 2023; Li et al., 2018; Recanatesi et al., 2019).

In language models, representational geometry has been used to analyze memorization, abstraction, and concept organization (Ethayarajh, 2019; Elhage et al., 2021). Dimensional collapse has also been associated with learning dynamics and feature specialization (Nakkiran et al., 2021). While prior work primarily considers intrinsic dimensionality as a function of task difficulty or training progress, we apply it to characterize epistemic regimes, showing that uncertain inputs consistently occupy higher-dimensional manifolds than factual ones across architectures and modalities.

Topological Analysis of Neural Representations. Topological data analysis (TDA), and persistent homology in particular, has been increasingly applied to study the global structure of neural representations (Guss & Salakhutdinov, 2018; Rieck et al., 2019). These methods have been used to identify class separability, phase transitions during training, and structural complexity of learned manifolds. Recent work has shown that topology can reveal differences between random and trained networks, as well as between robust and brittle representations (Hofer et al., 2017; Chen & Ong, 2022).

Decision Geometry and Confidence Dynamics. The geometry induced by softmax classifiers partitions representation space into Voronoi-like regions associated with vocabulary embeddings (Aurenhammer et al., 2021). Optimization under cross-entropy encourages representations to move away from decision boundaries and deeper into class basins, increasing margin (Soudry et al., 2018; Ji & Telgarsky, 2020). In deep networks, this dynamic has been linked to feature amplification and confidence calibration, with gradients implicitly driving representations toward simplex vertices.

Training Dynamics and Alignment Effects. Recent work has highlighted how representation structure evolves over training, with early layers forming general-purpose features and deeper layers specializing toward task objectives (Nakkiran et al., 2021). Studies of fine-tuning and alignment have further shown that post-training can substantially reshape internal representations and confidence calibration, often trading robustness for helpfulness or fluency.