TOKEN-FAR: HALLUCINATION DETECTION FROM A TOKEN LOCAL GEOMETRY PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucination remains a major obstacle to the reliable deployment of Large Language Models (LLMs). However, existing hallucination detection methods often rely on repeated generation, semantic comparison, or trained detectors, causing extra inference cost or requiring supervision. In this paper, we investigate whether the local geometry of the internal token representations from a single candidate response can provide an effective hallucination signal without the need for training. To this end, we propose **Token-FAR** (**Token**-level **F**actual **A**ffine **R**econstruction), a token-level geometric hallucination detector based on local affine reconstruction with respect to a factual reference dictionary. To be specific, in Token-FAR, we utilize the hidden states of factual answers to construct a factual reference dictionary, and then for each candidate content token, we compute an affine ridge reconstruction residual with respect to a local factual neighborhood. We observe that the content tokens in a candidate answer that are poorly reconstructed by the local factual neighborhood will result in larger residuals, which provide a signal to indicate an occurrence of hallucination. By aggregating the token-level residuals, we define a score for detecting answer-level hallucination. Our proposed Token-FAR requires only a single candidate response at inference time and does not need to train a parametric detector using hallucination labels. We conduct extensive experiments and ablation studies on four factuality benchmark datasets across two LLMs. Experimental results demonstrate the effectiveness of our proposed approach, showing that local affine reconstruction with respect to a factual reference dictionary can provide an effective geometric signal for hallucination detection.

1 INTRODUCTION

In nowadays, Large Language Models (LLMs) are deployed ubiquitous. However, LLMs can generate unsupported claims—which are often called hallucinations—as fluently as correct ones, making them difficult to detect from surface form alone (Ji et al., 2023; Huang et al., 2025). Hallucinations occur in question answering (Lin et al., 2022a; Li et al., 2023a), scientific factuality assessment (Chen et al., 2023), and knowledge-grounded dialogue (Dziri et al., 2022). We focus on a practical setting to detect hallucinations, i.e., scoring a single candidate response without requiring an external verifier or a classifier trained on hallucination labels.

Existing methods for hallucination detection only partly satisfy this requirement. For example, the existing likelihood-based scoring methods, e.g., (Malinin & Gales, 2021), are efficient, but they can be affected by lexical frequency, answer length, and surface uncertainty; Semantic-uncertainty methods, e.g., (Kuhn et al., 2023; Farquhar et al., 2024), and consistency-based methods, e.g., (Manakul et al., 2023), can better handle paraphrastic variation, but often require multiple sampled responses, which increases inference cost. Moreover, disagreement among responses does not necessarily indicate a factual error. Hidden states in LLMs provide an alternative source of evidence for hallucination detection. Prior works (Azaria & Mitchell, 2023; Burns et al., 2023; Du et al., 2024; Liu et al., 2026) have identified that the truthfulness-related signals in internal representations could help detect hallucinations. For example, Chen et al. (2024) measure semantic variation across multiple internal representations, retaining the repeated-generation cost of output-consistency methods; Hu et al. (2026) instead score a single response from projected hidden states, but rely on a parametric detector trained with hallucination labels. However, these results confirm the value of internal rep-

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.

resentations, but leave open whether a single response can be evaluated by directly measuring its support from factual representations without using hallucination label to train a detector.

In this paper, we attempt to address the hallucination detection problem from a token-level local geometry perspective. Specifically, for each candidate token, we measure how well its hidden representation can be reconstructed from nearby tokens which are extracted from factual answers. The tokens from factual answers provide a local reference support to interpret each candidate token and thus to define a hallucination scoring without modeling the hallucination distribution explicitly. This view is inspired by outlier detection approach based on self-representation, where inliers are assumed to lie in a union of low-dimensional subspaces (UoS) but outliers are deviating from UoS (You et al., 2017). We assume that, *at a token level, the hallucinations correspond to outlying tokens that are deviate from the factual context and the factual tokens correspond to low-dimensional semantic structure*[1] *due to complicated semantic constraints.*

We propose **Token-FAR** (Token-level Factual Affine Reconstruction), a token-level subspace reconstruction based scoring approach for hallucination detection. To be more specific, we construct a factual reference dictionary which consists of internal representations of factual tokens, and for each candidate content token, we compute a local affine ridge reconstruction with the nearby factual states in the dictionary, define a hallucination score via the reconstruction residual, and aggregate the token-level residuals to define an answer-level score. Our Token-FAR requires only one candidate response and does not need to train a parametric detector with hallucination labels.

Paper Contributions. Our contributions of the paper are highlighted as follows.

- We formulate the hallucination detection problem from a token level local geometry perspective, which connects hallucination detection problem to outlier detection task.
- We provide a geometric decomposition of the Token-FAR residual into off-subspace displacement and shrinkage along weakly supported directions within the local factual subspace, clarifying the geometric source of the hallucination signal.
- We conduct extensive experiments to evaluate the performance of our Token-FAR against nine baselines with two LLMs on four factuality benchmark datasets, verifying the efficacy and effectiveness of our proposed approach.

2 Related Work

2.1 Hallucination Detection from Output Uncertainty

Output-based hallucination detection methods often rely on model confidence or agreement across sampled responses. For example, Perplexity (Ren et al., 2023) and Token Entropy (Malinin & Gales, 2021) require little additional computation, but sequence-level uncertainty can also reflect benign lexical variation; semantic-uncertainty methods (Kuhn et al., 2023; Farquhar et al., 2024) mitigates this issue by grouping responses with equivalent meanings. Besides, Manakul et al. (2023) use disagreement across stochastic generations as a signal of uncertain knowledge; whereas Kadavath et al. (2022) study self-evaluation by asking language models to assess whether their own answers are correct. These approaches do not need to access to the hidden states, but the extra semantic comparison, repeated generation or self-evaluation would bring additional inference cost.

2.2 Truthfulness Signals in Internal Representations

Internal representations have been showing truthfulness-related signals. For instance, Azaria & Mitchell (2023) train classifiers on hidden activations to predict statement truthfulness; Burns et al. (2023) recover latent binary knowledge from unlabeled activations through logical consistency constraints; Marks & Tegmark (2023) identify linear truth-related directions that generalize across factual datasets; Li et al. (2023b) leverage truth-related activation directions for inference-time intervention.

[1]For example, the semantic low-dimensional structure is a union of low dimensional semantic manifolds (or semantic fiber manifolds).

Building on these observations, several methods use internal representations to construct hallucination signals or detectors. For example, Chen et al. (2024) derive EigenScore from the covariance spectrum of multiple responses; Du et al. (2024) train a classifier using automatically scored generations; Hu et al. (2026) apply a supervised detector after projecting token representations onto a proposed reasoning subspace. In contrast, our Token-FAR does not need to train a parametric detector with hallucination labels. Instead, it uses factual reference dictionary to compute a local reconstruction residual for each candidate token.

Truthfulness-related signals may also vary across model depth. For example, Chuang et al. (2024) exploit layer-wise differences by contrasting token distributions from different decoder layers. Several studies further report that intermediate-layer activations can be more informative than final-layer representations in some settings, e.g., (Azaria & Mitchell, 2023; Chen et al., 2024; Aiersilan, 2026). More recent methods model hidden-state variation across layers rather than relying on a single final-layer representation (Zhang et al., 2025; Kim et al., 2025). These observations motivate us to use a fixed architecture-relative layer.

2.3 Geometric and Subspace-Based Anomaly Detection

Geometric anomaly detection identifies atypical samples through representation-space structure. Feature-space distances, such as class-conditional Mahalanobis distance, have been widely used for out-of-distribution detection, e.g., (Lee et al., 2018). Recently, Liu et al. (2026) extend this perspective to hallucination detection using geometric relations among hidden representations, output weights, and decision boundaries. Our approach instead measures a factual support through local affine reconstruction from nearby factual hidden states. Subspace-based anomaly detection provides a closely relevant geometric perspective. An intriguing work is (You et al., 2017), which presents a self-representation-based outlier detection approach under a union of subspaces assumption. This work inspires us that the reconstruction residual can be used as a geometric signal of how well a representation is supported by the underlying data structure.

3 Method

In this section, we present our framework for hallucination detection, Token-FAR, which is illustrated in Figure 1. Section 3.1 defines the problem formulation, and Sections 3.2 introduce the construction of the factual reference dictionary. Sections 3.3-3.5 describe a local affine reconstruction, hallucination score, and answer-level score aggregation.

3.1 Problem Formulation

Let f_θ be an autoregressive language model with L decoder blocks. Given a question $\mathcal{X}$ and a candidate answer $\mathcal{Y} = (y_1, \ldots, y_n)$ consisting of n tokens, we need to define a scalar scoring function, which is denoted as $S(\mathcal{X}, \mathcal{Y})$, where a larger value of $S(\mathcal{X}, \mathcal{Y})$ indicates a higher likelihood of hallucination. Let g_ℓ denote the mapping from an input question $\mathcal{X}$ to the output of decoder block ℓ. We define $\boldsymbol{h}_i^{(\ell)} = g_\ell(\mathcal{X}, y_{\le i}) \in \mathbb{R}^d$ as the hidden state at the sequence position occupied by answer token y_i. Under causal masking, the representation $\boldsymbol{h}_i^{(\ell)}$ encodes y_i and its left context (i.e., $y_{<i}$) but not any future answer tokens (i.e., $y_{>i}$). We use the internal representation from the fixed architecture-specific layer ℓ_{mid}, where $\ell_{\text{mid}} := \lceil L/2 \rceil$. To be specific, the query representation $\boldsymbol{z}_i \in \mathbb{S}^{d-1}$ is obtained by L_2-normalized $\boldsymbol{h}_i^{(\ell_{\text{mid}})}$, where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$. We define the factual reference set $\mathcal{R}$ as the set of R question-answer pairs, i.e., $\mathcal{R} := \{(\mathcal{X}_m, \mathcal{Y}_m^\star)\}_{m=1}^R$, where $\mathcal{Y}_m^\star = (y_{m,1}^\star, \ldots, y_{m,T_m}^\star)$ denotes the factual answer token sequence associated with the reference question $\mathcal{X}_m$, and $T_m = |\mathcal{Y}_m^\star|$ denotes the number of tokens in $\mathcal{Y}_m^\star$.

3.2 Factual Representation Dictionary

For each reference pair $(\mathcal{X}_m, \mathcal{Y}_m^\star) \in \mathcal{R}$, we encode the complete question-answer sequence under teacher forcing and extract the normalized representation $\boldsymbol{z}_{m,j}^\star$ at the position of each factual answer

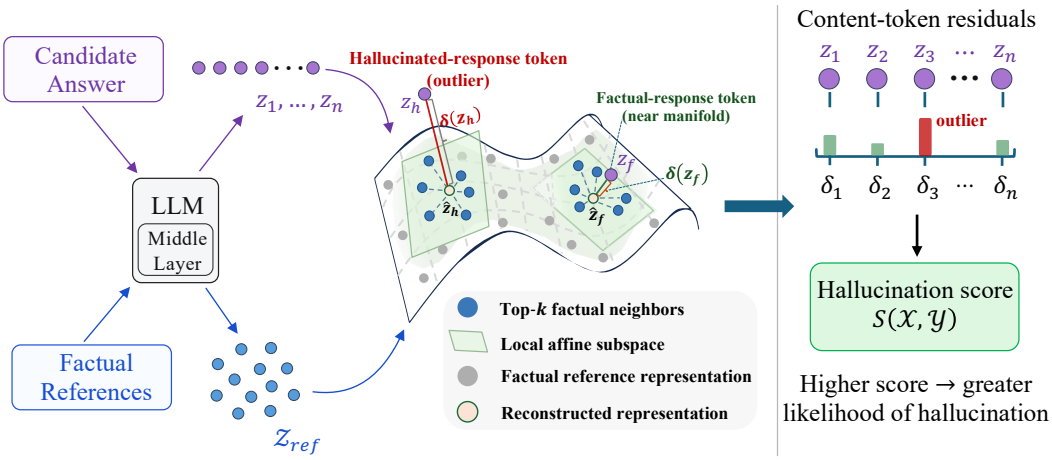


Figure 1: Overview of our proposed Token-FAR. Token-level representations for factual reference answers and candidate answers are extracted from a fixed middle layer. Each representation is taken at the token's own sequence position and is therefore conditioned on that token and its remaining context. The factual representations form the reference dictionary. For each candidate token representation $\boldsymbol{z}_i$, our Token-FAR retrieves a local factual neighborhood and computes an affine ridge reconstruction residual. The residuals over content tokens are then averaged to obtain an answer-level hallucination score, where a larger value indicate a higher likelihood of being hallucination.

token $\boldsymbol{y}^{\star}_{m,j}$. These representations form the token-level factual representation dictionary

$$\mathcal{Z}_{\text{ref}} = \bigcup_{m=1}^{R} \{\boldsymbol{z}^{\star}_{m,j} : 1 \le j \le T_m\} \subset \mathbb{S}^{d-1}. \tag{1}$$

For the normalized query representation $\boldsymbol{z} \in \mathbb{S}^{d-1}$, we find the top-$k$ reference representations with the highest cosine similarity from the factual reference dictionary $\mathcal{Z}_{\text{ref}}$, i.e.

$$\mathcal{N}_k(\boldsymbol{z}) = \underset{\substack{\mathcal{A} \subset \mathcal{Z}_{\text{ref}} \\ |\mathcal{A}|=k}}{\arg\max} \sum_{\boldsymbol{q} \in \mathcal{A}} \boldsymbol{z}^{\top} \boldsymbol{q}. \tag{2}$$

The neighborhood $\mathcal{N}_k(\boldsymbol{z}) = \{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k\}$ serves as the local factual dictionary used for building an affine reconstruction of $\boldsymbol{z}$.

3.3 Computing Local Affine Reconstruction

Local factual geometry. The factual reference dictionary $\mathcal{Z}_{\text{ref}}$ spans diverse lexical, syntactic, and semantic contexts. Rather than assuming that the factual token representations form a subspace, we instead assume they are spanning locally low-dimensional linear structures, which is motivated by the local linearity property used in manifold learning (Roweis & Saul, 2000). To be more precise, a sufficiently small neighborhood of a smooth manifold admits a local linear or affine approximation. Accordingly, for each candidate token $\boldsymbol{z}$, we characterize factual support through the geometry of the local factual neighborhood $\mathcal{N}_k(\boldsymbol{z})$. The candidate token representation associated with hallucination are more likely deviating from $\mathcal{N}_k(\boldsymbol{z})$ and thus are reconstructed with a larger residual.

Compute a local affine ridge reconstruction. Based on the local factual geometry, it is reasonable to assume that the factual neighborhood $\mathcal{N}_k(\boldsymbol{z})$ forms an affine subspace. Then we define a centralized factual dictionary as

$$\boldsymbol{D}_z = [\boldsymbol{q}_1 - \boldsymbol{\mu}_z, \quad \cdots, \quad \boldsymbol{q}_k - \boldsymbol{\mu}_z] \in \mathbb{R}^{d \times k}, \tag{3}$$

where $\boldsymbol{\mu}_z = \frac{1}{k}\sum_{j=1}^{k} \boldsymbol{q}_j$. It is apparent that the corresponding local affine subspace $\mathcal{A}_z$ can be represented by $\boldsymbol{\mu}_z + \text{col}(\boldsymbol{D}_z)$. Centralization separates the neighborhood location $\boldsymbol{\mu}_z$ from its local directions $\text{col}(\boldsymbol{D}_z)$, placing the query and its neighbors in a common local coordinate system.

Similarly, we centralize the query token $\boldsymbol{z}$ as $\widetilde{\boldsymbol{z}} = \boldsymbol{z} - \boldsymbol{\mu}_z$ and reconstruct it by solving

$$\boldsymbol{c}_z^\star = \arg\min_{\boldsymbol{c}\in\mathbb{R}^k} \|\widetilde{\boldsymbol{z}} - \boldsymbol{D}_z\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_2^2, \tag{4}$$

where $\lambda > 0$ is the tradeoff parameter. It is easy to check that rank deficiency arises because the centralized columns of $\boldsymbol{D}_z$ satisfy that $\boldsymbol{D}_z\mathbf{1} = \mathbf{0}$. The ℓ_2 norm regularization term makes the local reconstruction unique even when the columns in $\boldsymbol{D}_z$ are highly correlated. The unique closed-form solution is

$$\boldsymbol{c}_z^\star = (\boldsymbol{D}_z^\top\boldsymbol{D}_z + \lambda\boldsymbol{I}_k)^{-1}\boldsymbol{D}_z^\top\widetilde{\boldsymbol{z}}. \tag{5}$$

Given $\boldsymbol{c}_z^\star$, we have that the reconstructed representation for candidate token $\boldsymbol{z}$ is:

$$\widehat{\boldsymbol{z}} = \boldsymbol{\mu}_z + \boldsymbol{D}_z\boldsymbol{c}_z^\star. \tag{6}$$

Then we define a token-level reconstruction residual as

$$\delta(\boldsymbol{z}) = \|\boldsymbol{z} - \widehat{\boldsymbol{z}}\|_2, \tag{7}$$

which can be served as *a hallucination score*.

Although the optimization uses the columns in a centralized dictionary $\boldsymbol{D}_z$, we have that $\widehat{\boldsymbol{z}}$ is an affine combination of the original neighbors in $\mathcal{N}_k(\boldsymbol{z})$, with coefficients that sum to one. In Appendix B.1, we provide a brief justification.

Geometric interpretation for residual. The reconstruction residual has a direct geometric interpretation. Let $r = \mathrm{rank}(\boldsymbol{D}_z)$. Consider the compact SVD of the centralize local dictionary $\boldsymbol{D}_z = \boldsymbol{U}_r\boldsymbol{\Sigma}_r\boldsymbol{V}_r^\top$, where $\boldsymbol{\Sigma}_r = \mathrm{Diag}(\sigma_1, \ldots, \sigma_r)$ contains the nonzero singular values. Combining equation 5 and equation 6, we have that

$$\widehat{\boldsymbol{z}} = \boldsymbol{\mu}_z + \mathcal{P}_\lambda(\widetilde{\boldsymbol{z}}), \qquad \text{and} \quad \mathcal{P}_\lambda = \boldsymbol{U}_r\,\mathrm{Diag}\left(\frac{\sigma_1^2}{\sigma_1^2+\lambda}, \ldots, \frac{\sigma_r^2}{\sigma_r^2+\lambda}\right)\boldsymbol{U}_r^\top. \tag{8}$$

Denote $\widetilde{\boldsymbol{z}} = \boldsymbol{U}_r\boldsymbol{a} + \boldsymbol{b}$, where $\boldsymbol{a} = \boldsymbol{U}_r^\top\widetilde{\boldsymbol{z}}$ and $\boldsymbol{U}_r^\top\boldsymbol{b} = \mathbf{0}$. By the property of orthogonality, we have that

$$\delta(\boldsymbol{z})^2 = \underbrace{\|\boldsymbol{b}\|_2^2}_{\delta_\perp^2(\boldsymbol{z})} + \underbrace{\sum_{j=1}^{r}\left(\frac{\lambda}{\sigma_j^2+\lambda}\right)^2 a_j^2}_{\delta_\parallel^2(\boldsymbol{z})}, \tag{9}$$

where $\delta_\perp^2(\boldsymbol{z})$ is the orthogonal residual energy relative to the local factual subspace, and $\delta_\parallel^2(\boldsymbol{z})$ is the within-subspace residual induced by ridge shrinkage effect, with larger contributions along weakly supported directions. In experiments, we will provide a statistical distribution of the two components (See Figure 2) to interpret the rationale of using the reconstruction residual to detect hallucination.

In addition, it is easy to see that when $\lambda \to 0^+$, we have $\delta(\boldsymbol{z}) \to \mathrm{dist}(\boldsymbol{z}, \boldsymbol{\mu}_z + \mathrm{col}(\boldsymbol{D}_z))$ which is the Euclidean distance from candidate $\boldsymbol{z}$ to the local factual affine subspace. Thus, a larger residual indicates weaker support from the local factual neighborhood—implying hallucination occurs.

3.4 From Token Residuals to an Answer-Level Score

Since that the whitespace and punctuation tokens provide little semantic information and may bring tokenizer-specific spikes, we only score content tokens for answer-level hallucination detection and define an answer-level hallucination score by the averaged token-level residual, i.e.,

$$S(\mathcal{X}, \mathcal{Y}) = \frac{1}{|\mathcal{C}(\mathcal{Y})|}\sum_{i\in\mathcal{C}(\mathcal{Y})}\delta(\boldsymbol{z}_i), \tag{10}$$

where $\mathcal{C}(\mathcal{Y}) = \{i : \mathrm{decode}(y_i)$ contains at least one alphanumeric character$\}$ and if $\mathcal{C}(\mathcal{Y})$ is empty, we use all answer tokens instead. Note that taking an average over a token-level score makes the hallucination score comparable across the answers of different lengths and reduces the effect of isolated tokenization artifacts.

3.5 COMPUTATION COMPLEXITY

Our Token-FAR is a training-free hallucination detector. There is no need to train a parametric detector or use hallucination labels during scoring. The factual reference dictionary $\mathcal{Z}_{\text{ref}}$ is constructed once offline using token representations extracted from factual answers.

Let $N = |\mathcal{Z}_{\text{ref}}|$, M be the total number of candidate tokens, and d be the hidden dimension. For each query token, computing its cosine similarities to all N reference states costs $O(Nd)$, and selecting the top-k most similar token representations with a bounded min-heap costs $O(N \log k)$. In total, there is a time cost $O(M(Nd + N \log k))$ to construct the local factual dictionary for M candidate tokens. For each candidate token, constructing the local Gram matrix costs $O(dk^2)$, and solving the $k \times k$ ridge system costs $O(k^3)$. Thus the total time complexity for computing the local affine combination is $O\big(M\left(Nd + N \log k + dk^2 + k^3\right)\big)$.

4 EXPERIMENTS

To evaluate the performance of our proposed approach, we conduct experiments on four benchmark datasets with two open-weight instruction-tuned LLMs. We compare our proposed approach Token-FAR with single-response detectors and sampling-based detectors, and we investigate the effects of different reconstruction ways and different reference quality, followed by a sensitivity evaluation on the decoder depth.

4.1 EXPERIMENTAL SETUP

Datasets and LLMs We use four English hallucination detection benchmarks, including TruthfulQA (Lin et al., 2022a), HaluEval-QA (Li et al., 2023a), FELM-Science (Chen et al., 2023), and FaithDial (Dziri et al., 2022). These datasets cover open-domain falsehoods, knowledge-grounded QA, scientific factuality, and grounded dialogue. For FaithDial, we retain only pairs whose original response is explicitly annotated as hallucinated, excluding the edits that made only for cooperativeness or dialogue quality. Following (Liu et al., 2026), we evaluate two open-weight instruction-tuned LLMs: Qwen2.5-7B-Instruct (Yang et al., 2024) and Llama-3.2-3B-Instruct (Grattafiori et al., 2024). We use the released weights without further fine-tuning.

Evaluation Metrics We use AUROC (Area Under ROC curve) as the primary evaluation metric, in which AUROC measures how well a detector separates the factual responses from those hallucinated responses across different decision thresholds. Specifically, a higher AUROC indicates better discrimination, where 1.0 represents a perfect separation and 0.5 corresponds to a random guess.

Implementation Details. Implementation of this work is based on pytorch and transformers libraries. We use the normalized output of the middle decoder layer as the representation for hallucination detection. For candidate-response generation, we enable stochastic sampling with temperature of 0.5 and top-p of 0.99. A generated response is labeled "factual" if its maximum ROUGE-L score (Lin, 2004) exceeds 0.7 or its maximum semantic similarity exceeds 0.5. Semantic similarity is computed as the cosine similarity between sentence embeddings obtained from `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019). Across all experiments, we fix the number of local neighbors to $k = 10$ and set the ridge coefficient to $\lambda = 10^{-2}$.

4.2 MAIN RESULTS

To evaluate the effectiveness of our proposed approach, we compare Token-FAR to nine baseline methods that are training-free (i.e., not requiring to fit a parametric hallucination detector on labeled hallucination data), including Perplexity (Ren et al., 2023), Length-normalized Entropy (LN-Entropy) (Malinin & Gales, 2021), Energy score (Liu et al., 2020), P(True) (Kadavath et al., 2022), Lexical Similarity (Lin et al., 2022b), Semantic Entropy (Farquhar et al., 2024), EigenScore (Chen et al., 2024), NCI (Liu et al., 2026), and fDBD (Liu et al., 2026). Experimental results are reported in Table 1. From Table 1, we observe that our Token-FAR performs consistently well across datasets. To be specific, our Token-FAR achieves the best performance on seven cases across four datasets for two LLMs, except for dataset FELM-Science with Qwen2.5-7B-Instruct, in which Perplexity

Table 1: Hallucination detection performance across four datasets. All values are AUROC percentages. Best results within each model and dataset are in bold, and second-best results are underlined. "Single" indicates that no repeated candidate sampling is required.

Models	Methods	Single	TruthfulQA	HaluEval-QA	FELM-Science	FaithDial
Qwen2.5-7B-Instruct	Perplexity (Ren et al., 2023)	✓	46.3	69.4	**67.9**	66.9
	LN-Entropy (Malinin & Gales, 2021)	×	38.7	89.7	55.0	67.7
	Energy (Liu et al., 2020)	✓	<u>60.2</u>	67.7	53.6	<u>71.0</u>
	P(True) (Kadavath et al., 2022)	✓	58.7	<u>90.3</u>	59.3	70.2
	Lexical Similarity (Lin et al., 2022b)	×	58.9	15.3	53.6	52.4
	Semantic Entropy (Farquhar et al., 2024)	×	46.6	67.8	52.1	54.9
	EigenScore (Chen et al., 2024)	×	58.6	9.2	52.1	55.2
	NCI (Liu et al., 2026)	✓	51.7	61.5	51.9	69.1
	fDBD (Liu et al., 2026)	✓	50.8	55.7	54.1	68.9
	Token-FAR (Ours)	✓	**62.1**	**91.7**	<u>62.7</u>	**80.5**
Llama-3.2-3B-Instruct	Perplexity (Ren et al., 2023)	✓	64.2	88.2	63.0	64.5
	LN-Entropy (Malinin & Gales, 2021)	×	58.2	<u>95.0</u>	61.2	67.8
	Energy (Liu et al., 2020)	✓	69.4	57.3	62.5	68.1
	P(True) (Kadavath et al., 2022)	✓	64.1	85.1	<u>65.5</u>	<u>72.0</u>
	Lexical Similarity (Lin et al., 2022b)	×	67.8	55.0	57.0	50.1
	Semantic Entropy (Farquhar et al., 2024)	×	51.1	60.0	48.8	55.2
	EigenScore (Chen et al., 2024)	×	61.9	12.2	43.0	54.5
	NCI (Liu et al., 2026)	✓	<u>70.1</u>	66.9	64.0	68.4
	fDBD (Liu et al., 2026)	✓	68.3	76.8	63.9	65.8
	Token-FAR (Ours)	✓	**70.6**	**97.1**	**70.5**	**78.2**

(Ren et al., 2023) yields the best result. Its macro-average AUROC is 74.3% on Qwen2.5-7B and 79.1% on Llama-3.2-3B, compared to 69.6% and 71.7% for P(True) (Kadavath et al., 2022).

4.3 More Evaluation and Analysis

Evaluation on Reconstruction Geometry. To evaluate the effectiveness of using local affine ridge reconstruction, we conduct a set of experiments to compare our Token-FAR with two different ways to compute the answer-level hallucination score $S(\mathcal{X}, \mathcal{Y})$ defined in equation 10. To be specific, we use Local Linear ridge Reconstruction (LLR) to define the token residual $\delta(\boldsymbol{z}_i)$ in which $\boldsymbol{D}_z$ is not centralized, and we also use an average dissimilarity induced from cosine similarity to define the hallucination score $S(\mathcal{X}, \mathcal{Y})$, which is termed as Neighbor-based approach. Experimental results are shown in Table 2. We can read that the best performance is achieved when using local affine ridge reconstruction, which is implemented in our Token-FAR. Even using a local linear ridge reconstruction, the performance also beats the neighborhood based approach. These results suggest that the local affine ridge reconstruction does capture the special local geometry of the hallucinated answers with respect to the factual dictionary.

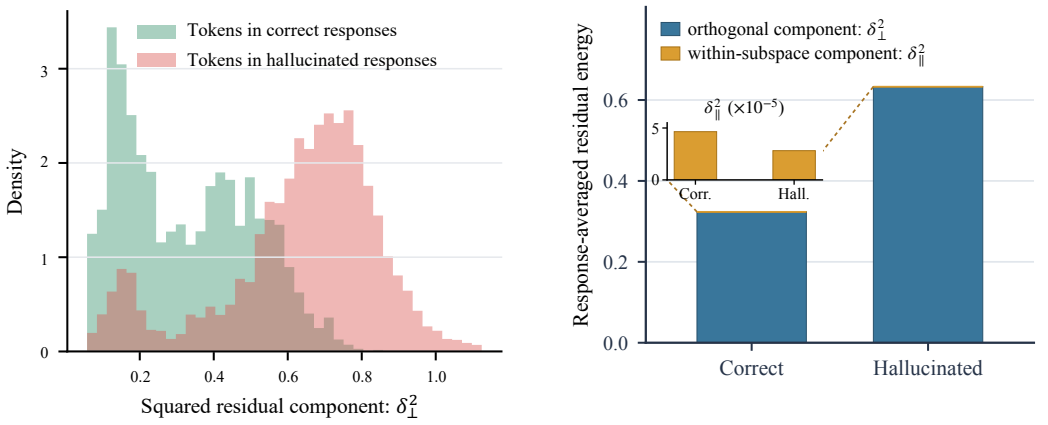
Table 2: Ablation study on different ways to define token residual.

Models	Methods	TruthfulQA	HaluEval-QA	FELM-Science	FaithDial
Qwen2.5-7B-Instruct	Neighbor	51.3	84.6	55.7	75.7
	Local Linear Ridge	<u>60.2</u>	<u>88.4</u>	<u>59.7</u>	<u>79.4</u>
	Local Affine Ridge	**62.1**	**91.7**	**62.7**	**80.5**
Llama-3.2-3B-Instruct	Neighbor	61.4	93.1	64.1	73.1
	Local Linear Ridge	<u>68.2</u>	<u>96.9</u>	<u>69.4</u>	<u>75.9</u>
	Local Affine Ridge	**70.6**	**97.1**	**70.5**	**78.2**

Evaluation on effect of reference-dictionary To evaluate the effect of the reference dictionary constructed in different ways, we conduct experiments on four datasets with both LLMs with respect to varied ways to construct the dictionary. To be specific, we construct a mixed dictionary, in which half of the factual reference tokens are replaced with tokens from incorrect answers, and a generated-response dictionary, in which the samples token representations are directly from the generated responses without using correctness labels. Experimental results are shown in Table 3. We observe that the detection performance is dramatically degenerated if a mixed dictionary or a generated-response reference dictionary is used. This confirms the importance of using a high-quality factual references for our Token-FAR.

Table 3: Ablation study on different ways to construct reference-dictionary.

Models	Methods	TruthfulQA	HaluEval-QA	FELM-Science	FaithDial
Qwen2.5-7B-Instruct	Mixed Dictionary	14.5	78.5	50.1	75.9
	Generated-Response Dictionary	43.5	70.6	46.1	72.6
	Factual Dictionary	**62.1**	**91.7**	**62.7**	**80.5**
Llama-3.2-3B-Instruct	Mixed Dictionary	16.2	85.7	61.5	74.0
	Generated-Response Dictionary	55.2	76.0	57.6	68.9
	Factual Dictionary	**70.6**	**97.1**	**70.5**	**78.2**



(a) Distribution of orthogonal residual energy

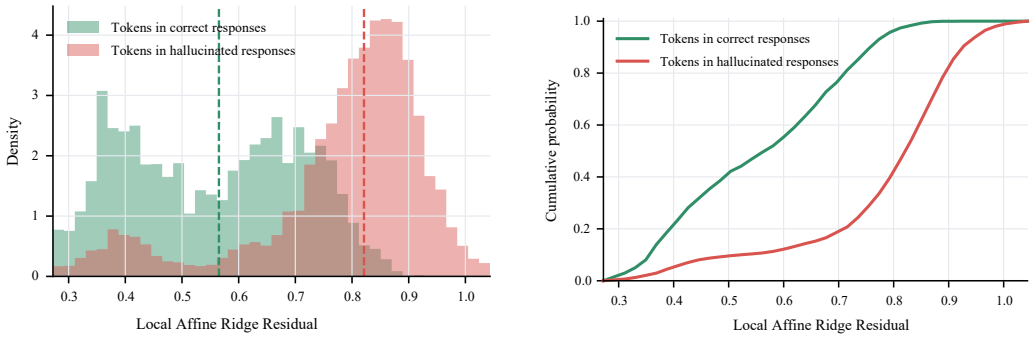
(b) Mean residual-energy decomposition

Figure 2: Residual decomposition for Llama-3.2-3B-Instruct on HaluEval-QA. **a)** Distribution of orthogonal residual energy $\delta_\perp^2$ for tokens from factual and hallucinated responses. **b)** Response-averaged mean residual energy decomposed into the orthogonal component $\delta_\perp^2$ and the within-subspace component $\delta_\parallel^2$ induced by ridge shrinkage.

Local factual geometry validation. To test whether the Token-FAR residual reflects the deviation from the local factual structure, we conduct experiments on HaluEval-QA dataset with Llama-3.2-3B-Instruct, in which we decompose the Token-FAR residual into an orthogonal component $\delta_\perp^2$ and a within-subspace component $\delta_\parallel^2$ (as defined in equation 9). We display the averaged $\delta_\perp^2$ and the averaged $\delta_\parallel^2$ for the hallucinated answers and correct answers, respectively, in Figure 2. From Figure 2, we observe that tokens from hallucinated responses exhibit substantially greater orthogonal residual energy than those from factual responses, whereas the within-subspace component induced by ridge shrinkage remains negligible. This indicates that the increased Token-FAR residual for hallucinated responses is driven primarily by greater deviation from the locally supported factual structure, providing empirical support for the geometric interpretation underlying our approach.

Comparison on residual distributions To further analyze the token-level reconstruction residuals, we conduct a set of experiments on HaluEval-QA with of Llama-3.2-3B, and display the distributions of reconstruction residuals in Figure 3. From Figure 3a, we observe that tokens from hallucinated responses tend to have larger residuals than those from correct responses. The empirical cumulative distributions in Figure 3b shows a dramatic gap between the correct responses and hallucinated responses, which provide a reliable evidence for response hallucination detection via our Token-FAR from a token-level local geometry perspective.

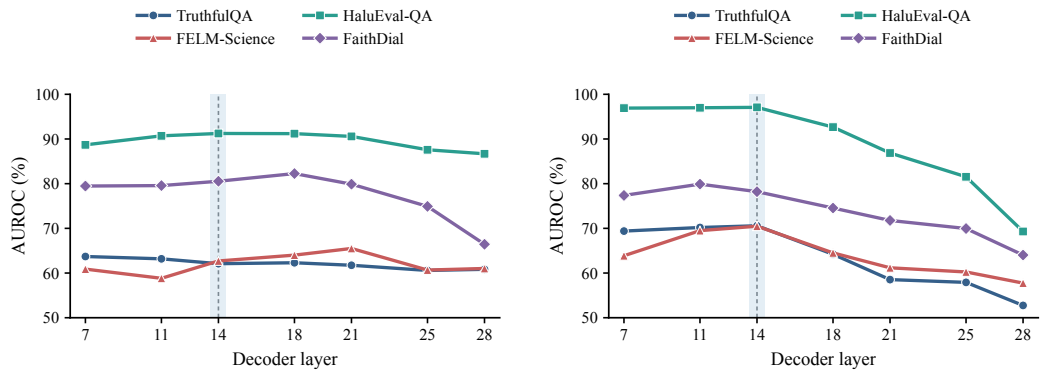
Layer sensitivity. Truthfulness signals can vary across decoder layers (Azaria & Mitchell, 2023; Chen et al., 2024; Aiersilan, 2026; Zhang et al., 2025). To evaluate the effect of representation depth on Token-FAR, we conduct experiments across different decoder layers on four datasets with both LLMs. Specifically, we evaluate $\ell \in \{7, 11, 14, 18, 21, 25, 28\}$. As shown in Figures 4, Qwen2.5-7B-Instruct remains relatively stable across intermediate layers over the four datasets,



(a) Residual density (b) Empirical cumulative distribution

Figure 3: Token-level reconstruction residual distributions on HaluEval-QA. **a)** Residual density. **b)** Empirical cumulative distribution.

whereas Llama-3.2-3B-Instruct generally achieves its strongest performance in the intermediate region and declines toward later layers. These results support the use of representations from a fixed architecture-specific middle layer, which provides a stable and informative representation for Token-FAR across both LLMs and datasets.



(a) (b)

Figure 4: Effect of decoder layer on hallucination detection performance. **(a)** Qwen2.5-7B-Instruct. **(b)** Llama-3.2-3B-Instruct.

5 Conclusion

We have presented a simple yet principled approach for single-response hallucination detection based on local affine reconstruction from factual references, called Token-FAR, which evaluates candidate token representations according to their support from local factual geometry. We conducted extensive experiments on four datasets with two LLMs and experimental results demonstrated superior performance across the two LLMs and four benchmark datasets without the need to train a hallucination-labeled detector.

Limitations Token-FAR measures local geometric support rather than factual correctness; hallucination-associated representations that remain close to the factual affine structure may therefore receive small residuals. Its performance also depends on the quality and coverage of the factual reference dictionary.

Ethics statement

This study analyzes existing factuality benchmarks and open-weight language models; it does not involve recruiting participants or collecting new human-subject data. Our work fully adheres to research integrity principles, and no ethical concerns have arisen during the course of this study.

Reproducibility statement

We provide details on dataset construction, reference–evaluation separation, candidate labeling, and leakage control in Appendix A. Section 4.1 describes the evaluated models, sampling settings, representation layer, and reconstruction hyperparameters. For each model–dataset pair, all methods are evaluated on the same fixed set of candidate responses.

AI use statement

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript, including sentence rephrasing, grammar checking, and improving readability. All AI-assisted text was reviewed by the authors. The authors take full responsibility for the final content of this work and ensure that all LLM-assisted text complies with relevant ethical standards and does not involve plagiarism or research misconduct.

References

Aizierjiang Aiersilan. Hallucination is linearly decodable from mid-layer hidden states in quantized LLMs. *arXiv preprint arXiv:2606.02628*, 2026.

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.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023.

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*, 2024.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. FELM: Benchmarking factuality evaluation of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. HaloScope: Harnessing unlabeled LLM generations for hallucination detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 102948–102972, 2024.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490, 2022.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Junjie Hu, Gang Tu, ShengYu Cheng, Jinxin Li, Jinting Wang, Rui Chen, Zhilong Zhou, and Dongbo Shan. HARP: Hallucination detection via reasoning subspace projection. In *The Fourteenth International Conference on Learning Representations*, 2026.

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):42:1–42:55, 2025.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Hazel Kim, Tom A. Lamb, Adel Bibi, Philip Torr, and Yarin Gal. Detecting LLM hallucination through layer-wise information deficiency: Analysis of ambiguous prompts and unanswerable questions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32310–32322, 2025.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, 2023a.

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

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004.

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*, pp. 3214–3252, 2022a.

Zi Lin, Jeremiah Zhe Liu, and Jingbo Shang. Towards collaborative neural-symbolic graph semantic parsing via uncertainty. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 4160–4173, 2022b.

Litian Liu, Reza Pourreza, Yubing Jian, Yao Qin, and Roland Memisevic. From out-of-distribution detection to hallucination detection: A geometric view. *arXiv preprint arXiv:2602.07253*, 2026.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475, 2020.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *The Ninth International Conference on Learning Representations*, 2021.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *NeurIPS Workshop on Attributing Model Behavior at Scale*, 2023.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. 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. 3982–3992. Association for Computational Linguistics, 2019.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Chong You, Daniel P. Robinson, and René Vidal. Provable self-representation based outlier detection in a union of subspaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3395–3404, 2017.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. ICR probe: Tracking hidden state dynamics for reliable hallucination detection in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 17986–18002, 2025.

A Implementation and Reproducibility Details

A.1 Dataset Construction

Table 4 summarizes the dataset-specific construction. For each dataset, examples are shuffled with a fixed random seed before the reference and evaluation subsets are selected. Reference questions and evaluation questions are disjoint, and no answer associated with an evaluation question is used to construct the factual representation dictionary. The dictionary is built by teacher-forcing one factual answer for each reference question.

For HaluEval-QA, FELM-Science, and FaithDial, candidate answers and binary factuality labels are provided by the corresponding benchmark. For TruthfulQA, candidate answers are generated by the evaluated model and labeled factual when either the maximum ROUGE-L score exceeds 0.7 or the maximum cosine similarity between `all-MiniLM-L6-v2` sentence embeddings exceeds 0.5.

Table 4: Dataset construction used in the experiments. "Ref." and "Eval." denote the numbers of reference and evaluation questions, respectively.

Dataset	Configuration	Ref.	Eval.	Candidate source
TruthfulQA	generation	300	500	model generation
HaluEval-QA	qa_samples	300	1,000	dataset answers
FELM-Science	science	50	1,000	dataset responses
FaithDial	plain_text	300	1,000	paired dataset answers

For HaluEval-QA, the prompt contains the question and supplied knowledge when available, and the benchmark-provided factual and hallucinated answers form the paired candidates. For FELM-Science, a response is treated as factual only when all annotated response segments are factual. For FaithDial, we retain only examples whose edit annotations explicitly identify hallucination; the human-edited faithful response and the original hallucinated response are evaluated under the same knowledge and dialogue history. FaithDial prompts are truncated to at most 768 tokens.

A.2 Evaluation Protocol and Leakage Control

All methods for a given model–dataset pair are evaluated on the same fixed candidate set. Evaluation labels are never used to construct the factual representation dictionary, retrieve local neighbors, or fit the local reconstruction, and they are accessed only for computing AUROC. Because reference and evaluation questions are disjoint, no evaluation answer contributes token representations to the factual dictionary. Higher scores consistently indicate a greater likelihood of hallucination.

B Detailed Derivations

This section provides detailed derivations for the reference-local affine reconstruction introduced in Section 3.3.

B.1 Equivalent affine coefficients

Because the columns of $\boldsymbol{D}_z$ are centered around their local mean, they satisfy a zero-sum constraint. From Equation 3,

$$\boldsymbol{D}_z\mathbf{1} = \sum_{j=1}^{k} (\boldsymbol{q}_j - \boldsymbol{\mu}_z) = \sum_{j=1}^{k} \boldsymbol{q}_j - k\left(\frac{1}{k}\sum_{j=1}^{k} \boldsymbol{q}_j\right) = \mathbf{0}. \tag{11}$$

Hence, $\mathbf{1}$ lies in the right null space of $\boldsymbol{D}_z$, implying $\mathrm{rank}(\boldsymbol{D}_z) \leq k-1$. Thus, the centered local dictionary is necessarily rank-deficient, as noted in Section 3.3.

The reconstruction is $\widehat{\boldsymbol{z}} = \boldsymbol{\mu}_z + \boldsymbol{D}_z \boldsymbol{c}_z^\star$. Writing the neighborhood matrix as $\boldsymbol{Z}_z = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k]$ and noting that $\boldsymbol{D}_z = \boldsymbol{Z}_z - \boldsymbol{\mu}_z \mathbf{1}^\top$ yields

$$\widehat{\boldsymbol{z}} = \boldsymbol{\mu}_z + (\boldsymbol{Z}_z - \boldsymbol{\mu}_z \mathbf{1}^\top)\boldsymbol{c}_z^\star \tag{12}$$

$$= \boldsymbol{Z}_z \boldsymbol{c}_z^\star + \boldsymbol{\mu}_z(1 - \mathbf{1}^\top \boldsymbol{c}_z^\star) \tag{13}$$

$$= \boldsymbol{Z}_z \left[\boldsymbol{c}_z^\star + \frac{1 - \mathbf{1}^\top \boldsymbol{c}_z^\star}{k} \mathbf{1} \right]. \tag{14}$$

Thus the equivalent coefficient vector

$$\boldsymbol{\alpha} = \boldsymbol{c}_z^\star + \frac{1 - \mathbf{1}^\top \boldsymbol{c}_z^\star}{k} \mathbf{1} \tag{15}$$

satisfies $\mathbf{1}^\top \boldsymbol{\alpha} = 1$. The centered ridge reconstruction is equivalent to an affine combination of the original local neighbors. Under a common translation of the query and its neighborhood, the reconstruction shifts by the same amount while the reconstruction residual remains unchanged. The resulting geometry therefore depends on relative local structure rather than on the choice of origin.

B.2 Spectral decomposition of the ridge residual

Let $r = \operatorname{rank}(\boldsymbol{D}_z)$ and consider the compact singular value decomposition

$$\boldsymbol{D}_z = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r^\top, \qquad \boldsymbol{\Sigma}_r = \operatorname{Diag}(\sigma_1, \ldots, \sigma_r), \tag{16}$$

where every σ_j is positive. The columns of $\boldsymbol{U}_r$ and $\boldsymbol{V}_r$ are orthonormal bases for the column and row spaces of $\boldsymbol{D}_z$, respectively. From the ridge solution, the centered reconstruction can be written as follows:

$$\widehat{\boldsymbol{z}} - \boldsymbol{\mu}_z = \boldsymbol{D}_z (\boldsymbol{D}_z^\top \boldsymbol{D}_z + \lambda \boldsymbol{I}_k)^{-1} \boldsymbol{D}_z^\top \widetilde{\boldsymbol{z}}. \tag{17}$$

Because $\boldsymbol{D}_z^\top \boldsymbol{D}_z = \boldsymbol{V}_r \boldsymbol{\Sigma}_r^2 \boldsymbol{V}_r^\top$, its regularized inverse separates into the row space of $\boldsymbol{D}_z$ and its orthogonal complement:

$$\begin{aligned}(\boldsymbol{D}_z^\top \boldsymbol{D}_z + \lambda \boldsymbol{I}_k)^{-1} = \boldsymbol{V}_r (\boldsymbol{\Sigma}_r^2 + \lambda \boldsymbol{I}_r)^{-1} \boldsymbol{V}_r^\top \\ + \lambda^{-1} (\boldsymbol{I}_k - \boldsymbol{V}_r \boldsymbol{V}_r^\top).\end{aligned} \tag{18}$$

The second term in Equation 18 vanishes after left multiplication by $\boldsymbol{D}_z$ because $\boldsymbol{V}_r^\top (\boldsymbol{I}_k - \boldsymbol{V}_r \boldsymbol{V}_r^\top) = 0$. Substituting Equations 16 and 18 into Equation 17 gives

$$\widehat{\boldsymbol{z}} - \boldsymbol{\mu}_z = \boldsymbol{U}_r \boldsymbol{\Sigma}_r (\boldsymbol{\Sigma}_r^2 + \lambda \boldsymbol{I}_r)^{-1} \boldsymbol{\Sigma}_r \boldsymbol{U}_r^\top \widetilde{\boldsymbol{z}} \tag{19}$$

$$= \boldsymbol{U}_r \operatorname{Diag}\left(\frac{\sigma_1^2}{\sigma_1^2 + \lambda}, \ldots, \frac{\sigma_r^2}{\sigma_r^2 + \lambda} \right) \boldsymbol{U}_r^\top \widetilde{\boldsymbol{z}} = \mathcal{P}_\lambda \widetilde{\boldsymbol{z}}. \tag{20}$$

Thus, $\mathcal{P}_\lambda$ is a regularized projection onto $\operatorname{col}(\boldsymbol{D}_z)$. Each singular direction is retained with gain $\sigma_j^2/(\sigma_j^2 + \lambda)$ rather than with the unit gain of an orthogonal projector.

To obtain the residual, decompose the centered query into components parallel and orthogonal to the local column space:

$$\widetilde{\boldsymbol{z}} = \boldsymbol{U}_r \boldsymbol{a} + \boldsymbol{b}, \qquad \boldsymbol{a} = \boldsymbol{U}_r^\top \widetilde{\boldsymbol{z}}, \qquad \boldsymbol{b} = (\boldsymbol{I}_d - \boldsymbol{U}_r \boldsymbol{U}_r^\top)\widetilde{\boldsymbol{z}}, \qquad \boldsymbol{U}_r^\top \boldsymbol{b} = \mathbf{0}. \tag{21}$$

Let $\boldsymbol{\Gamma}_\lambda = \operatorname{Diag}(\sigma_1^2/(\sigma_1^2 + \lambda), \ldots, \sigma_r^2/(\sigma_r^2 + \lambda))$. Using Equation 20, the residual vector becomes

$$\boldsymbol{z} - \widehat{\boldsymbol{z}} = \widetilde{\boldsymbol{z}} - \mathcal{P}_\lambda \widetilde{\boldsymbol{z}} \tag{22}$$

$$= \boldsymbol{U}_r \boldsymbol{a} + \boldsymbol{b} - \boldsymbol{U}_r \boldsymbol{\Gamma}_\lambda \boldsymbol{a} \tag{23}$$

$$= \boldsymbol{b} + \boldsymbol{U}_r \operatorname{Diag}\left(\frac{\lambda}{\sigma_1^2 + \lambda}, \ldots, \frac{\lambda}{\sigma_r^2 + \lambda} \right) \boldsymbol{a}. \tag{24}$$

The two terms in Equation 24 are orthogonal. Moreover, $\boldsymbol{U}_r$ preserves Euclidean norms on its coefficient space. Applying the Pythagorean theorem yields

$$\delta(\boldsymbol{z})^2 = \|\boldsymbol{z} - \widehat{\boldsymbol{z}}\|_2^2 \tag{25}$$

$$= \|\boldsymbol{b}\|_2^2 + \left\| \mathrm{Diag}\left(\frac{\lambda}{\sigma_1^2 + \lambda}, \ldots, \frac{\lambda}{\sigma_r^2 + \lambda} \right) \boldsymbol{a} \right\|_2^2 \tag{26}$$

$$= \|\boldsymbol{b}\|_2^2 + \sum_{j=1}^{r} \left(\frac{\lambda}{\sigma_j^2 + \lambda} \right)^2 a_j^2. \tag{27}$$

The first term is the squared displacement outside the local factual subspace. The second captures regularization-induced shrinkage within the subspace. Directions with small σ_j receive stronger shrinkage because they are weakly supported by the local factual neighbors.

Finally, define the local affine set $\mathcal{A}_z = \boldsymbol{\mu}_z + \mathrm{col}(\boldsymbol{D}_z)$. The orthogonal projection of $\boldsymbol{z}$ onto $\mathcal{A}_z$ is $\boldsymbol{\mu}_z + \boldsymbol{U}_r \boldsymbol{U}_r^\top \widetilde{\boldsymbol{z}}$, so

$$\mathrm{dist}(\boldsymbol{z}, \mathcal{A}_z) = \|(\boldsymbol{I}_d - \boldsymbol{U}_r \boldsymbol{U}_r^\top)\widetilde{\boldsymbol{z}}\|_2 = \|\boldsymbol{b}\|_2. \tag{28}$$

For every nonzero singular value, $\lambda/(\sigma_j^2 + \lambda) \to 0$ as $\lambda \to 0^+$. Equation 27 then implies

$$\lim_{\lambda \to 0^+} \delta(\boldsymbol{z}) = \mathrm{dist}(\boldsymbol{z}, \boldsymbol{\mu}_z + \mathrm{col}(\boldsymbol{D}_z)) \,. \tag{29}$$

For finite λ, the residual is bounded by

$$\mathrm{dist}(\boldsymbol{z}, \mathcal{A}_z)^2 \le \delta(\boldsymbol{z})^2 \le \|\boldsymbol{z} - \boldsymbol{\mu}_z\|_2^2. \tag{30}$$

The lower bound isolates off-subspace displacement, while the additional ridge term preserves sensitivity to directions that are present but weakly supported in the local dictionary.

C Case Studies

C.1 Hallucination Detection Cases

We compare Token-FAR against the baseline methods on HaluEval-QA using Llama-3.2-3B-Instruct. The optimal classification thresholds is determined by maximizing the G-Mean value, which is defined as G-Mean $= \sqrt{\mathrm{TPR} \times (1\text{-FPR})}$. The resulting thresholds are 2.2064 for Perplexity, 2.0810 for LN-Entropy, 0.0601 for EigenScore, and 0.7067 for Token-FAR. Scores at or above these thresholds indicate hallucination. For LexicalSimilarity, a score larger than the threshold indicates non-hallucination. For other metrics, a score smaller than the thresholds indicates non-hallucination. ✓(✗) indicates that the hallucination is (not) correctly identified by the metric. Green and red answers denote factual and hallucinated responses, respectively.

Challenging detection case. The supplied fact dates Sherry Thomas's award to 2011, whereas the response changes only the year to 2012. All four baselines place this fluent, minimally altered answer on the factual side of their thresholds. Token-FAR is the only method that correctly detects the hallucination. This case shows that a localized factual substitution can preserve low sequence-level uncertainty and high lexical agreement while departing from the reference-local token geometry.

Question: When did Sherry Thomas receive the Romance Writers of America RITA Award for Best Historical Romance?
Reference fact: The supplied knowledge states that *His at Night* won the award in 2011.
LLM answer: Sherry Thomas received the Romance Writers of America RITA Award for Best Historical Romance in 2012.
Response label: Hallucinated

Perplexity:	2.0642 ✗
LN-Entropy:	1.2319 ✗
LexicalSimilarity:	0.1407 ✗
EigenScore:	0.0382 ✗
Token-FAR:	0.7683 ✓

Representative detection cases. The following six cases contain three hallucinated and three factual responses. All five methods correctly classify the Amman–Riyadh and Cristian Chivu examples. The four remaining examples produce both correct and incorrect baseline decisions, while Token-FAR classifies every response correctly.

Question: Tooheys Extra Dry and Hahn Brewery were accquired by which company in 2009?
Reference fact: The supplied knowledge says the Lion beverages group was acquired by the Japanese Kirin Company.
LLM answer: Tooheys Extra Dry and Hahn Brewery were accquired by Lion beverages group in 2009.
Response label: Hallucinated

Perplexity:	2.3433 ✓
LN-Entropy:	3.7682 ✓
LexicalSimilarity:	0.1322 ✗
EigenScore:	0.0385 ✗
Token-FAR:	0.7904 ✓

Question: Lugton and Dunlop are both villages located in what country?
Reference fact: The supplied knowledge locates both Lugton and Dunlop in East Ayrshire, Scotland.
LLM answer: Lugton and Dunlop are both located in England.
Response label: Hallucinated

Perplexity:	5.6974 ✓
LN-Entropy:	1.8875 ✗
LexicalSimilarity:	0.1088 ✓
EigenScore:	0.0722 ✓
Token-FAR:	0.7667 ✓

Question: In which capital city is the Arab Academy for Banking and Financial Sciences located?
Reference fact: The supplied knowledge locates the university in Amman, Jordan.
LLM answer: The Jordan banking and financial sciences academy is located in Riyadh.
Response label: Hallucinated

Perplexity:	7.5815 ✓
LN-Entropy:	4.8626 ✓
LexicalSimilarity:	0.0981 ✓
EigenScore:	0.0655 ✓
Token-FAR:	0.7864 ✓

Question: What cultural region in County Galway, Ireland, is where Irish writer of thrillers and television screenplays, Manning O'Brine, was born?
Reference fact: The supplied knowledge places his birth in Connemara, County Galway.
LLM answer: Connemara
Response label: Factual

Perplexity:	1.0283 ✓
LN-Entropy:	0.2995 ✓
LexicalSimilarity:	0.1818 ✓
EigenScore:	0.0923 ✗
Token-FAR:	0.6056 ✓

Question: Mircea Chivu Stadium is named after the father of which Romanian soccer player?
Reference fact: Mircea Chivu was the father of Romanian soccer player Cristian Chivu.
LLM answer: Cristian Chivu
Response label: Factual

Perplexity:	1.0417 ✓
LN-Entropy:	0.6219 ✓
LexicalSimilarity:	0.2250 ✓
EigenScore:	0.0593 ✓
Token-FAR:	0.6426 ✓

Question: Who directed a film that included Sarah Manninen?
Reference fact: Sarah Manninen appeared in *The Prince and Me*, directed by Martha Coolidge.
LLM answer: Martha Coolidge
Response label: Factual

Perplexity:	1.0473 ✓
LN-Entropy:	0.5466 ✓
LexicalSimilarity:	0.1302 ✗
EigenScore:	0.0687 ✗
Token-FAR:	0.5605 ✓

Failure cases. The first response reverses an explicit chronological comparison, yet all five methods classify it as factual. The second response is a one-token answer that incorrectly changes the correct count from five to two. Although its post-token representation directly encodes the erroneous token "two", it remains well supported by the retrieved factual neighborhood, resulting in a Token-FAR score below the hallucination threshold. LN-Entropy also misses this error, whereas Perplexity, LexicalSimilarity, and EigenScore detect it. Together, these cases show that local factual support is not a perfect surrogate for factual correctness: a hallucination-associated representation may remain close to the local factual structure and therefore yield a small reconstruction residual.

Question: Which game was published first, *Villa Paletti* or *Taj Mahal*?
Reference fact: *Taj Mahal* was published in 2000, one year before *Villa Paletti*.
LLM answer: Villa Paletti was published first.
Response label: Hallucinated

Perplexity:	1.9008 ✗
LN-Entropy:	1.9168 ✗
LexicalSimilarity:	0.2161 ✗
EigenScore:	0.0312 ✗
Token-FAR:	0.6631 ✗

Question: Heatstroke stars an actress who has written how many non-fiction books?
Reference fact: The supplied knowledge states that Danica McKellar wrote five non-fiction books.
LLM answer: two
Response label: Hallucinated

Perplexity:	401.0443 ✓
LN-Entropy:	1.5039 ✗
LexicalSimilarity:	0.0001 ✓
EigenScore:	0.0965 ✓
Token-FAR:	0.3812 ✗

C.2 Token-Level Residual Visualization

Figure 5 visualizes the global percentile of each token-level reconstruction residual for six hallucinated responses from Llama-3.2-3B-Instruct on HaluEval-QA. In the five multi-token responses, the answer-level scores range from 0.737 to 0.833, and the largest residuals are concentrated in particular content spans rather than distributed uniformly across the response. This pattern illustrates that

Token-FAR produces non-uniform token-level signals, with some representations receiving much weaker support from the local factual geometry than others. In contrast, the one-token answer "two" has both a low residual percentile and a low answer-level score of 0.381. Although its post-token representation directly encodes the erroneous token, it remains locally well supported by the retrieved factual neighborhood and therefore yields a small reconstruction residual.

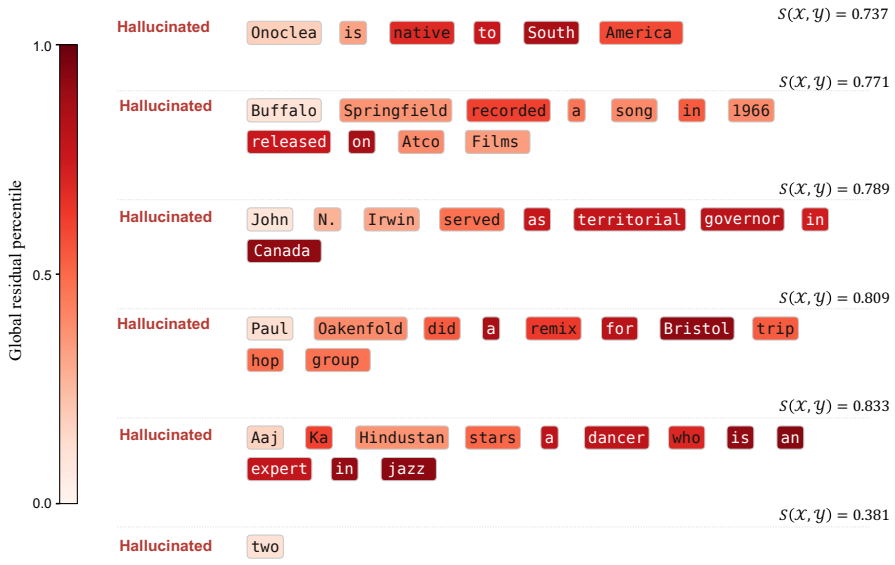


Figure 5: Token-level residual visualization for six hallucinated Llama-3.2-3B-Instruct responses on HaluEval-QA. Each cell represents a content token; darker red indicates a higher global percentile of its local affine reconstruction residual. $S(\mathcal{X}, \mathcal{Y})$ denotes the mean residual over the selected answer tokens. The final row shows a one-token failure case in which the erroneous token remains well supported by the retrieved factual neighborhood.

D Additional Analysis

D.1 Sensitivity to the Number of Neighbors

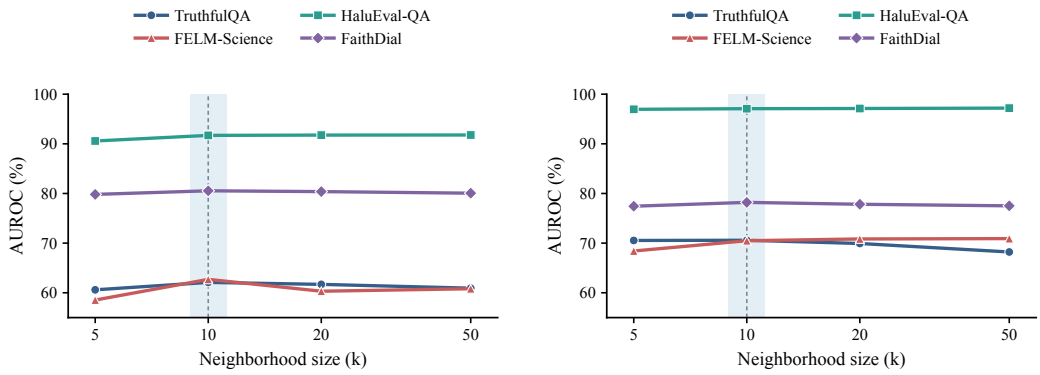
We vary the number of retrieved factual neighbors over $k \in \{5, 10, 20, 50\}$. As shown in Figure 6, Token-FAR remains stable across a broad range of neighborhood sizes, particularly on HaluEval-QA and FaithDial. TruthfulQA and FELM-Science show slightly larger variation, with the most noticeable drop occurring for Qwen at $k = 5$.

D.2 Sensitivity to the Ridge Coefficient

As shown in Figure 7, Token-FAR is largely insensitive to λ across the tested range. Performance on HaluEval-QA and FaithDial remains nearly unchanged, while TruthfulQA and FELM-Science exhibit only modest variation.

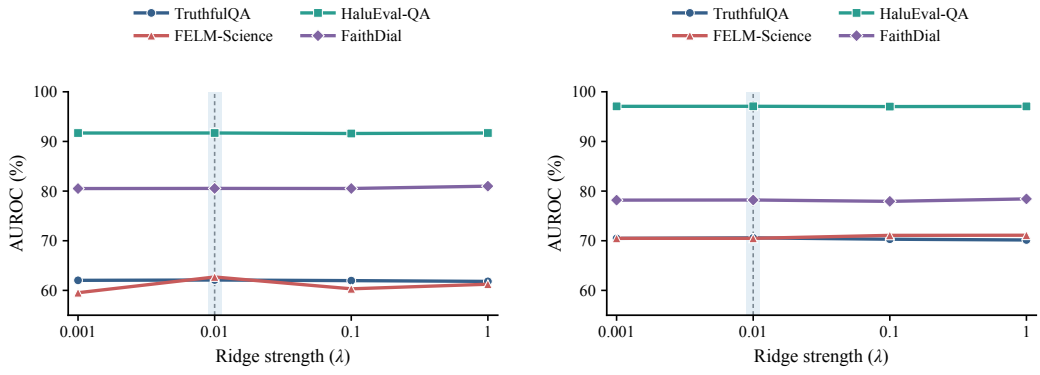
D.3 Sensitivity to the Factual Reference Set Size

We further examine the sensitivity of Token-FAR to the size of the factual reference set using Llama-3.2-3B-Instruct on HaluEval-QA. As shown in Figure 8, with only 25 factual examples, Token-FAR achieves 96.73% AUROC, only 0.35 percentage points below the 300-example setting. Increasing the reference set to 100 examples raises AUROC to 97.05%, within 0.03 points of the full setting. Performance then largely saturates, and the slight increase at 200 examples falls within the observed



(a) Qwen2.5-7B-Instruct (b) Llama-3.2-3B-Instruct

Figure 6: Effect of neighborhood size k on hallucination detection performance.



(a) Qwen2.5-7B-Instruct (b) Llama-3.2-3B-Instruct

Figure 7: Effect of the ridge coefficient λ on hallucination detection performance.

variation across runs. These results suggest that Token-FAR can establish effective local factual geometry from a relatively small reference set and does not require a large reference dictionary to achieve stable performance.

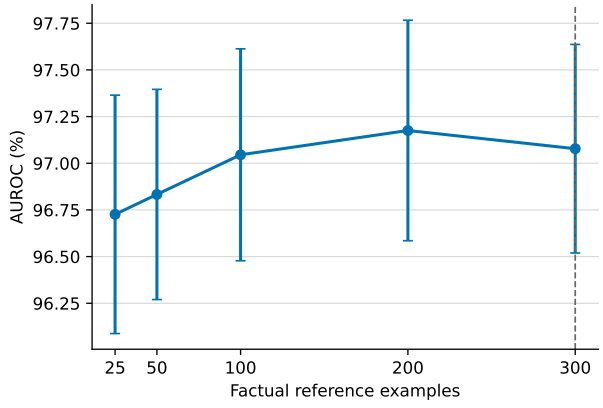


Figure 8: Effect of factual reference-set size on Token-FAR performance on HaluEval-QA with Llama-3.2-3B-Instruct. Points denote mean AUROC, and error bars indicate 95% uncertainty intervals.