GRA: GROUNDING RATIO ANALYSIS FOR PROMPT-ANCHORED UPDATE TRAJECTORIES IN HALLUCINATION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) can generate fluent but unsupported or factually incorrect responses, undermining their reliability in real-world deployment. Existing white-box hallucination detectors often focus on hidden-state snapshots or aggregate internal representations, leaving the generation-time update process underexplored. In this paper, we shift the focus from static hidden-state analysis to update anchoring and propose **Grounding Ratio Analysis (GRA)**, a white-box hallucination detector that measures prompt anchoring in generation-time residual updates. For each generated token and Transformer layer, GRA measures whether a residual-stream update remains anchored to the prompt region or shifts toward the generated-prefix region, forming a token-layer grounding trajectory. We summarize this trajectory with mean-trend pooling and use a lightweight probe for response-level reliability estimation. Experiments on five open-weight LLMs and five question-answering and hallucination detection datasets show that GRA achieves the highest average AUROC among evaluated methods on every evaluated model. Further ablations show that temporal trajectory information is important, prompt-anchoring signals appear across different update sources, and raw GRA trajectory statistics are already directly separable. These results suggest that prompt-anchored update dynamics provide an effective and interpretable signal for hallucination detection.

1 INTRODUCTION

Large language models (LLMs) have rapidly become general-purpose systems for natural language interaction, showing strong performance in question answering, instruction following, reasoning, and knowledge-intensive tasks Zhao et al. (2023); Minaee et al. (2024). Their ability to produce fluent and contextually coherent responses has led to broad adoption in open-ended user-facing applications. As users increasingly rely on LLM-generated answers as actionable knowledge, it becomes important to determine whether a generated response should be trusted.

A central obstacle to reliable deployment is hallucination: LLMs can generate responses that are fluent but unsupported or factually incorrect Huang et al. (2025). Such failures are especially problematic when users cannot easily verify every answer, as the surface fluency of an answer may obscure the absence of factual support. Hallucination detection therefore aims to estimate the reliability of generated responses and identify outputs that are likely to be untrustworthy.

Existing hallucination detection methods estimate reliability from different available signals. Output-side methods use confidence scores, likelihoods, semantic consistency, or repeated generations, while white-box methods exploit hidden-state features, activation subspaces, or learned readouts Xia et al. (2025). Although these approaches are effective, many of them summarize generation as either an output-side score or a static internal representation. This abstraction misses an important aspect of autoregressive decoding: a response is not produced in one step, but through a sequence of residual-stream updates written across generated tokens and layers.

This limitation motivates a generation-time view of hallucination detection. Recent work has shown that hidden-state dynamics can provide useful reliability signals Zhang et al. (2025), suggesting that hallucination-related evidence may be distributed across generated tokens and layers. However,

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.

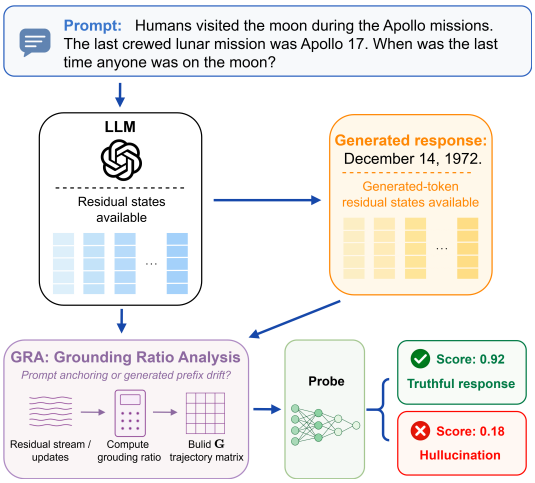

Figure 1: Intuition of GRA. GRA builds a token-layer grounding trajectory $\mathbf{G}$ from residual-stream update alignments and feeds it to a lightweight probe for hallucination detection.

existing dynamic analyses mainly characterize generic representation changes or component contributions, without explicitly describing what these updates are grounded in during generation. This leaves open how to measure whether generation-time updates remain anchored to the prompt or drift toward self-conditioned continuation.

To address this gap, we propose **Grounding Ratio Analysis (GRA)**, a white-box hallucination detector that measures prompt anchoring in generation-time residual updates, as illustrated in Figure 1. For each generated token and Transformer layer, GRA compares the absolute projection energy of a residual update onto the prompt region and the generated-prefix region. This produces a token-layer grounding trajectory that describes how prompt anchoring evolves during decoding. We then summarize the trajectory with mean-trend pooling and use a lightweight probe to produce a response-level reliability score.

Experiments across five open-weight LLMs and five question-answering and hallucination detection datasets show that GRA achieves the highest average AUROC among the evaluated methods on each evaluated model. Additional analyses support the role of prompt-anchored update trajectories: temporal pooling is important, prompt-anchoring signals are observed across multiple residual update sources, and raw GRA trajectory statistics already contain hallucination-relevant information before supervised probe training. These results suggest that GRA benefits from a meaningful update-anchoring structure in the generation process, rather than from probe capacity alone. Our contributions are summarized as follows:

- We introduce Grounding Ratio Analysis (GRA), a white-box hallucination detection framework that measures whether generation-time residual updates remain anchored to the prompt or shift toward the generated prefix.
- We construct token-layer grounding trajectories and show that mean-trend pooling provides an effective representation for response-level hallucination detection.
- We evaluate GRA across five open-weight LLMs and five datasets, showing that it outperforms strong baselines and that prompt-anchored update dynamics provide a robust hallucination-detection signal.

2 Related Work

Output-Side Uncertainty and Consistency Detection. A common line of hallucination detection estimates reliability from output-side signals, such as token likelihood, predictive entropy Malinin & Gales (2020), perplexity Ren et al. (2022), lexical similarity Lin et al. (2022b), verbalized confidence Yin et al. (2023); Xiong et al. (2024), and semantic consistency or semantic uncertainty Kuhn

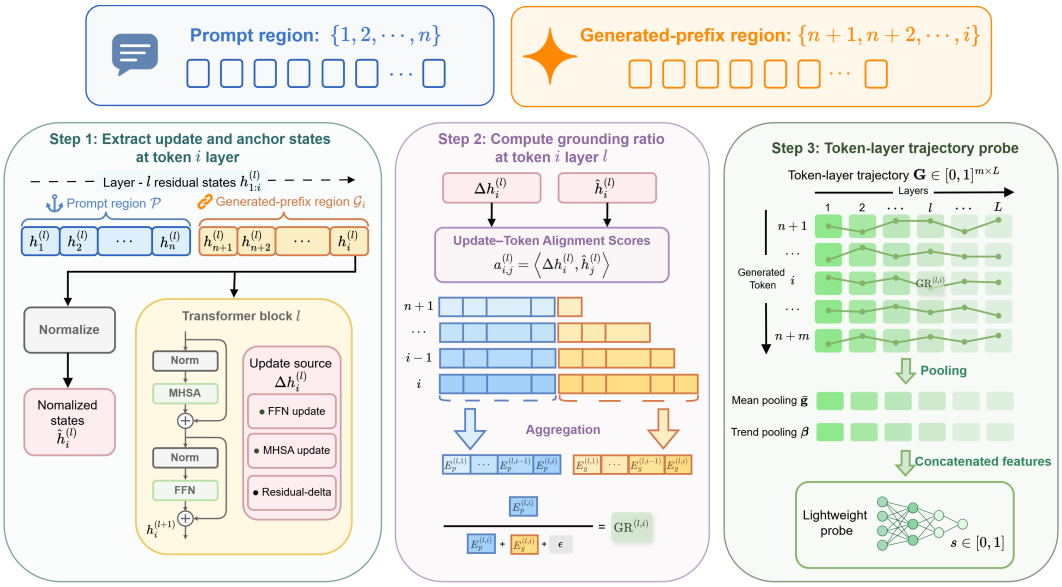

Figure 2: Overview of GRA. At each generated token and Transformer layer, GRA compares residual-update alignment with the prompt and generated prefix to form a token-layer grounding trajectory. Mean-trend pooling and a lightweight probe yield a response-level hallucination score.

et al. (2023); Farquhar et al. (2024). Likelihood- and entropy-based methods use token probabilities to measure uncertainty during decoding, while consistency-based methods generate multiple responses and treat disagreement as evidence of uncertainty or hallucination Manakul et al. (2023). These methods are broadly applicable, but they rely on output-side behavior rather than directly inspecting the residual-stream updates through which hidden representations evolve during generation. GRA complements this line by using white-box generation-time update signals from a single generated response.

White-Box Representation and Subspace Methods. White-box hallucination detection methods exploit internal activations of LLMs, based on the observation that model internals encode signals related to truthfulness or hallucination. INSIDE/EigenScore measures semantic dispersion in hidden-state embedding space from sampled responses Chen et al. (2024), while latent-knowledge and truthfulness-direction methods show that activation space can reveal whether a model is likely to be truthful or untruthful Burns et al. (2022); Azaria & Mitchell (2023). HaloScope identifies hallucination-related activation subspaces from unlabeled generations and trains a truthfulness classifier using projection-based membership estimation Du et al. (2024). LLM-Check further studies multiple internal and output-side signals, including hidden states, attention maps, and prediction probabilities Sriramanan et al. (2024). These methods demonstrate that LLM activations contain rich reliability signals, but they mainly analyze static hidden-state geometry, activation snapshots, sentence-level embeddings, or subspace membership, making it difficult to characterize how residual-stream updates evolve during generation. GRA builds on this white-box perspective but shifts the unit of analysis from static activation geometry to generation-time update anchoring.

Probe-Based Detection and Residual Dynamics. Probe-based methods provide a lightweight way to read task-relevant information from intermediate representations, and structural probing has shown that linguistic structure can be recovered from hidden representations Hewitt & Manning (2019). Recent hallucination detection work further uses probes to exploit reliability signals in hidden activations for token- or entity-level detection Obeso et al. (2025); Liang & Wang (2025). However, probes over hidden states often treat internal states as features to be decoded, without explicitly characterizing how reliability signals arise during generation. ICR Probe moves beyond static hidden-state probing by tracking hidden-state dynamics and component contributions to residual-stream updates Zhang et al. (2025). GRA also studies generation dynamics, but measures where updates are anchored: the prompt or the generated prefix.

3 METHOD

3.1 PROBLEM FORMULATION

We consider an L-layer causal language model f. Given a prompt sequence $\mathbf{x}^{\mathrm{p}} = (x_1, \ldots, x_n)$, the model generates a response $\mathbf{x}^{\mathrm{g}} = (x_{n+1}, \ldots, x_{n+m})$ autoregressively. The prompt sequence is generic, encompassing instructions, questions, or context passages when available. For each generated position $i \in \{n+1, \ldots, n+m\}$, the token x_i is produced from $P_f(\cdot \mid x_{<i})$ under the chosen decoding strategy, where $x_{<i}$ denotes the prefix before position i.

We study hallucination detection as response-level reliability estimation: given $(\mathbf{x}^{\mathrm{p}}, \mathbf{x}^{\mathrm{g}})$, the detector assigns a scalar score indicating whether the response is likely to be grounded in the prompt sequence rather than hallucinated. We focus on the white-box setting, where residual-stream states and layer-wise update vectors are accessible. Let d denote the hidden dimension. Let $h_i^{(l)} \in \mathbb{R}^d$ denote the residual-stream state of token x_i before the l-th Transformer block, and let $\Delta h_i^{(l)} \in \mathbb{R}^d$ denote a layer-wise update produced within that block for a generated token x_i. In our main instantiation, $\Delta h_i^{(l)} = \Delta h_{\mathrm{ffn},i}^{(l)}$, where $\Delta h_{\mathrm{ffn},i}^{(l)}$ denotes the feed-forward network (FFN) sublayer update written to the residual stream.

3.2 GROUNDING RATIO ANALYSIS

For each generated position i, GRA measures the relative alignment of the update $\Delta h_i^{(l)}$ with two causally visible regions: the prompt region $\mathcal{P} = \{1, \ldots, n\}$ and the generated-prefix region $\mathcal{G}_i = \{n+1, \ldots, i\}$. The latter includes the current token but excludes future generated tokens.

For each layer l, we normalize the residual-stream states:

$$\hat{h}_j^{(l)} = \frac{h_j^{(l)}}{\|h_j^{(l)}\|_2 + \epsilon}, \quad j \in \{1, \ldots, i\} \tag{1}$$

where ϵ is a small constant for numerical stability.

For a fixed generated position i and layer l, we first compute the alignment between the current update and each causally visible residual direction:

$$a_{i,j}^{(l)} = \left\langle \Delta h_i^{(l)}, \hat{h}_j^{(l)} \right\rangle = \sum_{r=1}^{d} \Delta h_{i,r}^{(l)} \hat{h}_{j,r}^{(l)}, \quad j \in \{1, \ldots, i\} \tag{2}$$

Here $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner product over the hidden dimension. Since $\hat{h}_j^{(l)}$ is normalized, $a_{i,j}^{(l)}$ is the projection coefficient of the update $\Delta h_i^{(l)}$ along the residual-state direction of token x_j.

We then compute the projection energy of $\Delta h_i^{(l)}$ over the prompt and generated-prefix regions by summing absolute alignment magnitudes:

$$E_{\mathrm{p}}^{(l,i)} = \sum_{j \in \mathcal{P}} \left| a_{i,j}^{(l)} \right| \tag{3}$$

$$E_{\mathrm{g}}^{(l,i)} = \sum_{j \in \mathcal{G}_i} \left| a_{i,j}^{(l)} \right| \tag{4}$$

where $E_{\mathrm{p}}^{(l,i)}$ and $E_{\mathrm{g}}^{(l,i)}$ denote the projection energy assigned to the prompt region and generated-prefix region, respectively.

The grounding ratio is defined as:

$$\mathrm{GR}^{(l,i)} = \frac{E_{\mathrm{p}}^{(l,i)}}{E_{\mathrm{p}}^{(l,i)} + E_{\mathrm{g}}^{(l,i)} + \epsilon} \tag{5}$$

By construction, $\mathrm{GR}^{(l,i)} \in [0, 1]$. Larger values indicate stronger relative prompt anchoring, while smaller values indicate a shift toward the generated-prefix region. GRA should be interpreted as an

internal update-anchoring signal rather than a direct factuality measure: it does not verify facts by itself, but measures whether the model's update remains more prompt-conditioned or shifts toward self-conditioned continuation. We therefore aggregate these local signals across generated tokens and layers into a token-layer grounding trajectory.

3.3 Token-Layer Grounding Trajectory

The grounding ratio $\mathrm{GR}^{(l,i)}$ provides a local signal for a specific generated token and a specific layer. To estimate reliability at the response level, GRA collects these local signals across the entire generation process. For each generated position $i \in \{n+1, \ldots, n+m\}$, we define its layer-wise grounding vector as:

$$\mathbf{g}_i = \left[\mathrm{GR}^{(1,i)}, \ldots, \mathrm{GR}^{(L,i)}\right] \in [0,1]^L \tag{6}$$

Stacking these vectors over all generated positions yields the token-layer grounding trajectory:

$$\mathbf{G} = \begin{bmatrix} \mathbf{g}_{n+1} \\ \mathbf{g}_{n+2} \\ \vdots \\ \mathbf{g}_{n+m} \end{bmatrix} \in [0,1]^{m \times L} \tag{7}$$

Each row of $\mathbf{G}$ corresponds to one generated token, and each column corresponds to one Transformer layer. Thus, $\mathbf{G}$ captures how prompt anchoring evolves over both decoding time and network depth. Since response lengths vary across examples, we next aggregate this trajectory into a fixed-dimensional representation.

3.4 Temporal Pooling and Probe

Since response lengths vary, we pool the trajectory $\mathbf{G} \in [0,1]^{m \times L}$ into a fixed-dimensional feature. We use two complementary statistics over the generated-token dimension.

Mean pooling measures the average prompt anchoring at each layer:

$$\bar{\mathbf{g}} = \frac{1}{m} \sum_{i=n+1}^{n+m} \mathbf{g}_i \in \mathbb{R}^L \tag{8}$$

Trend pooling measures the temporal change of prompt anchoring at each layer. Let $z_i = (i - \mu)/(\sigma + \epsilon)$ be the normalized generated position for $i \in \{n+1, \ldots, n+m\}$, where μ and σ are the mean and standard deviation of the index set $\{n+1, \ldots, n+m\}$. We compute:

$$\boldsymbol{\beta} = \frac{\sum_{i=n+1}^{n+m} z_i \mathbf{g}_i}{\sum_{i=n+1}^{n+m} z_i^2 + \epsilon} \in \mathbb{R}^L \tag{9}$$

where each entry of $\boldsymbol{\beta}$ is the linear trend of prompt anchoring over generated-token positions for one layer.

The final GRA feature concatenates both statistics:

$$\mathbf{f}_{\mathrm{GRA}} = [\bar{\mathbf{g}}; \boldsymbol{\beta}] \in \mathbb{R}^{2L} \tag{10}$$

Mean pooling captures the overall prompt anchoring level, while trend pooling captures how prompt anchoring changes along generated-token positions. We use their concatenation, referred to as mean-trend pooling, as the default GRA representation and validate this choice in Section 5.1.

Finally, a lightweight MLP probe h_θ maps $\mathbf{f}_{\mathrm{GRA}}$ to a bounded reliability score:

$$s = \sigma\left(h_\theta(\mathbf{f}_{\mathrm{GRA}})\right) \in [0,1] \tag{11}$$

where σ is the sigmoid function, and s predicts the response-level reliability. During training, the probe is optimized with binary cross-entropy using automatically constructed correctness labels. The probe architecture and training details are provided in Appendix C.

Model	Method	SQuAD	NQ-Open	TriviaQA	TruthfulQA	HaluEval	Avg.
Llama-2-7B	Perplexity	64.41	65.15	64.51	62.21	56.91	62.64
	LN-Entropy	50.75	50.40	50.43	66.08	50.51	53.63
	EigenScore	55.42	50.65	50.78	54.41	50.20	52.29
	HaloScope	72.00	84.66	52.79	80.91	72.59	72.59
	ICR Probe	73.67	88.99	80.84	80.10	81.29	80.98
	GRA (Ours)	**74.15**	**89.35**	**82.00**	**88.13**	**81.62**	**83.05**
Mistral-7B	Perplexity	54.46	75.98	50.46	**79.62**	51.48	62.40
	LN-Entropy	51.17	52.70	54.38	54.92	54.73	53.58
	EigenScore	50.71	54.06	53.20	51.51	56.70	53.24
	HaloScope	50.81	64.62	70.07	78.49	51.19	63.03
	ICR Probe	67.82	72.44	87.70	77.15	91.08	79.24
	GRA (Ours)	**71.44**	**92.14**	**88.96**	77.26	**94.72**	**84.90**
Llama-3-8B	Perplexity	61.56	55.74	69.29	78.24	58.44	64.66
	LN-Entropy	53.47	55.49	51.77	56.93	53.18	54.17
	EigenScore	52.84	53.93	50.87	54.11	51.54	52.66
	HaloScope	81.18	84.61	63.35	**78.63**	80.19	77.59
	ICR Probe	83.07	87.22	79.05	66.86	86.60	80.56
	GRA (Ours)	**84.64**	**87.30**	**79.67**	76.17	**88.10**	**83.18**
Qwen3-8B	Perplexity	50.20	55.75	51.29	76.86	56.88	58.20
	LN-Entropy	50.15	56.39	51.36	50.65	50.73	51.86
	EigenScore	50.54	56.75	50.45	52.46	52.71	52.58
	HaloScope	73.17	85.43	67.08	74.83	50.09	70.12
	ICR Probe	81.48	96.38	88.90	75.25	71.60	82.72
	GRA (Ours)	**81.61**	**97.50**	**89.24**	**79.46**	**87.34**	**87.03**
Qwen3-14B	Perplexity	56.13	59.87	52.64	**84.61**	69.34	64.52
	LN-Entropy	50.74	56.12	50.08	56.96	51.62	53.10
	EigenScore	51.28	52.55	50.01	52.52	50.08	51.29
	HaloScope	50.28	83.64	50.13	59.05	50.17	58.65
	ICR Probe	73.67	96.30	93.60	65.87	**94.29**	84.75
	GRA (Ours)	**76.40**	**97.56**	**93.80**	79.93	94.01	**88.34**

Table 1: Main experimental results for response-level hallucination detection, evaluated by AUROC (%). The best performance within each model block is highlighted in **bold**, and the second best is underlined. GRA achieves the highest average AUROC across all five evaluated models.

3.5 Update Source Variants

GRA supports multiple update sources. By default, FFN-GRA uses $\Delta h_i^{(l)} = \Delta h_{\mathrm{ffn},i}^{(l)}$, the update written by the FFN sublayer into the residual stream.

We also consider Attn-GRA, which uses the MHSA update $\Delta h_i^{(l)} = \Delta h_{\mathrm{attn},i}^{(l)}$. Residual-delta-GRA uses the full block-level residual change $\Delta h_i^{(l)} = h_i^{(l+1)} - h_i^{(l)}$. Here, $\Delta h_{\mathrm{attn},i}^{(l)}$ is the self-attention update, and $h_i^{(l)}$, $h_i^{(l+1)}$ are the residual-stream states before and after block l.

These variants test whether prompt anchoring extends beyond FFN updates. Section 5.2 compares the three choices; FFN-GRA is the default because it directly captures a feed-forward residual-stream update.

4 Experiments

4.1 Experimental Setup

Models and datasets. We evaluate five open-weight LLMs—**Llama-2-7B-Chat** Touvron et al. (2023), **Mistral-7B-Instruct-v0.3** Jiang et al. (2023), **Llama-3-8B-Instruct** Grattafiori et al. (2024), **Qwen3-8B**, and **Qwen3-14B** Yang et al. (2025)—on five QA and hallucination datasets:

SQuAD Rajpurkar et al. (2018), **NQ-Open** Kwiatkowski et al. (2019), **TriviaQA** Joshi et al. (2017), **TruthfulQA** Lin et al. (2022a), and **HaluEval** Li et al. (2023).

Evaluation Metric. We report the area under the receiver operating characteristic curve (**AUROC**) as the primary metric. AUROC is threshold-independent and measures the probability that a randomly chosen correct response receives a higher reliability score than a randomly chosen incorrect response. For each dataset, we report AUROC on a 20% held-out test split stratified by answer correctness.

Correctness Measurement. Following prior work Chen et al. (2024); Du et al. (2024); Ma et al. (2026), we construct automatic correctness labels using the maximum ROUGE-L score between the generated response and reference answers. A response is labeled correct when this score exceeds 0.5. We further test label robustness under ROUGE-L > 0.7 and BLEURT > 0.5.

Baselines. We compare GRA with both training-free and training-based hallucination detection baselines. Training-free baselines include **Perplexity** Ren et al. (2022), **LN-Entropy** Malinin & Gales (2020), and **EigenScore** Chen et al. (2024). Training-based baselines include **HaloScope** Du et al. (2024) and **ICR Probe** Zhang et al. (2025). For each model–dataset pair, all training-based methods are trained only on the (80%) training split and evaluated on the same (20%) held-out evaluation split. Details of these baselines are in Appendix A.

4.2 Main Results

Table 1 presents the main AUROC results across five open-weight LLMs and five evaluation datasets. GRA achieves the highest average AUROC on all evaluated models, showing that token-layer grounding trajectories provide a useful reliability signal across different model families and scales rather than being tied to a single backbone.

Compared with the strongest training-based baseline, ICR Probe, GRA improves the average AUROC on every evaluated model, with an average gain of 3.65 AUROC points across the five models. GRA also improves over HaloScope in average AUROC across all models, suggesting that explicitly measuring prompt anchoring of internal updates captures information that is not fully reflected by subspace-based representation analysis.

At the dataset level, the improvements are especially clear on SQuAD, NQ-Open, and TriviaQA, where GRA obtains the best AUROC for every evaluated model. GRA also performs strongly on HaluEval, achieving the best result on four of the five models and remaining competitive on Qwen3-14B. Representative gains include HaluEval on Qwen3-8B, where AUROC increases from 71.60% with ICR Probe to 87.34%, and NQ-Open on Mistral-7B, where AUROC increases from 72.44% to 92.14%. Across related model variants, including both Llama-family models and the two Qwen3 scales, GRA maintains the strongest average performance, suggesting that prompt-anchored update trajectories provide a reusable signal across multiple open-weight LLMs. These results are consistent with the intuition that hallucinated responses often involve a shift away from prompt-conditioned updates toward self-conditioned continuation over the generated prefix.

Robustness to response length. We further examine whether GRA mainly exploits response length rather than grounding dynamics. Since GRA aggregates token-layer trajectories over generated positions, response length is a natural potential confound. We group responses into short, medium, and long length ranges. As shown in Table 2, GRA remains effective across all three length ranges and across model families, suggesting that its advantage is not merely driven by answer length.

Label robustness. We examine whether the results are sensitive to the automatically constructed correctness labels by re-evaluating the saved ICR Probe and GRA predictions under alternative label protocols. Besides the default ROUGE-L > 0.5 criterion, we consider a stricter ROUGE-L > 0.7 criterion and a BLEURT > 0.5 semantic-similarity criterion. As shown in Table 3, GRA consistently outperforms ICR Probe across all evaluated models and label protocols, suggesting that its advantage is not tied to a particular labeling threshold or scoring function.

Model	Short	Medium	Long
Llama-2-7B	66.59	72.03	87.38
Mistral-7B	74.65	76.16	79.50
Llama-3-8B	69.60	78.88	75.35
Qwen3-8B	68.66	77.28	76.44
Qwen3-14B	74.96	74.65	75.77

Table 2: Answer-length robustness of GRA. Values are average AUROC (%).

Model	**ROUGE-L** > 0.5		**ROUGE-L** > 0.7		**BLEURT** > 0.5	
	ICR Probe	GRA	ICR Probe	GRA	ICR Probe	GRA
Llama-2-7B	80.98	**83.05**	81.47	**83.97**	79.91	**81.87**
Mistral-7B	79.24	**84.90**	87.58	**92.16**	80.64	**85.26**
Llama-3-8B	80.56	**83.18**	81.11	**85.30**	80.12	**82.55**
Qwen3-8B	82.72	**87.03**	84.90	**90.49**	82.34	**88.09**
Qwen3-14B	84.75	**88.34**	89.75	**93.20**	87.24	**89.47**
Avg.	81.65	**85.30**	84.96	**89.02**	82.05	**85.45**

Table 3: Label robustness under alternative protocols. Values are average AUROC (%).

5 Ablation and Analysis

We conduct ablations on Llama-3-8B and Qwen3-14B across the five datasets in Section 4, and evaluate update-source variants on all five models because their behavior may be architecture-dependent. These analyses are designed to test whether GRA's gains come from the proposed token-layer trajectory structure, the choice of residual update source, or the learned probe aggregation. Unless otherwise specified, probe-based ablations report five-seed average AUROC for more robust comparison under training variance, with full results in Appendix E.

5.1 Temporal Pooling Ablation

We first study how to pool $\mathbf{G} \in [0,1]^{m \times L}$ over generated-token positions into a response-level feature. Table 4**(a)** compares five strategies on the two representative models: First, Last, Mean, Trend, and Mean-Trend. These strategies summarize the generated-token dimension while preserving layer-wise grounding information.

First-token pooling performs substantially worse than the other strategies on both models, suggesting that early generated-token positions do not fully expose hallucination-related grounding signals. Last, Mean, and Trend are much stronger, confirming the value of aggregating the trajectory over generated tokens. Mean-Trend achieves the strongest average performance on both representative models and is best or near-best across datasets, supporting our choice to jointly model the average strength of prompt anchoring and its temporal drift over the response.

5.2 Effect of Update Sources

We next examine whether GRA depends on the choice of residual update source. Following Section 3.5, we compare Attn-GRA, FFN-GRA, and Residual-delta-GRA using the same Mean-Trend pooling and probe architecture. Table 4**(b)** reports the average AUROC over the five datasets for each evaluated model.

All three update sources achieve competitive performance, suggesting that prompt-anchoring information is present across different residual-stream updates rather than being an artifact of a single module. At the same time, no update source consistently dominates across models: Residual-delta-GRA is strongest on the two Llama models, while Attn-GRA is strongest on Mistral-7B and both Qwen3 models. FFN-GRA remains close to the best source on each model. We therefore use FFN-GRA as the default instantiation of GRA, not because it is uniformly the strongest variant, but

(a) Temporal pooling

Model	Pooling	SQuAD	NQ-Open	TriviaQA	TruthfulQA	HaluEval
Llama-3-8B	First	67.06	76.19	67.46	59.84	69.02
	Last	82.34	88.39	79.56	78.04	86.44
	Mean	82.74	88.21	78.18	78.48	87.02
	Trend	82.82	88.67	76.82	77.80	86.62
	Mean-Trend	**84.76**	**88.92**	**79.86**	**78.73**	**88.66**
Qwen3-14B	First	60.43	80.98	83.12	63.12	85.54
	Last	73.37	94.85	93.61	77.76	94.51
	Mean	74.96	94.20	93.83	77.12	94.51
	Trend	**76.88**	95.10	93.71	75.55	94.33
	Mean-Trend	76.66	**95.21**	**94.12**	**78.10**	**94.84**

(b) Update sources

Model	Attn	FFN	Residual
Llama-2-7B	82.17	82.57	82.58
Mistral-7B	85.78	84.59	84.38
Llama-3-8B	83.53	84.19	84.73
Qwen3-8B	87.35	86.44	86.47
Qwen3-14B	88.16	87.79	87.68

(c) Direct separability

Model	Statistic	SQuAD	NQ	Trivia	TQA	Halu
Llama-3-8B	Mean	75.93	87.00	73.56	57.94	82.46
	Trend	76.72	86.58	71.19	59.93	81.68
	Full GRA	**84.76**	**88.92**	**79.86**	**78.73**	**88.66**
Qwen3-14B	Mean	70.43	91.90	92.80	78.81	93.46
	Trend	70.96	92.54	92.97	**78.85**	93.67
	Full GRA	**76.66**	**95.21**	**94.12**	78.10	**94.84**

Table 4: Ablation analyses: **(a)** temporal pooling, **(b)** update-source ablation, and **(c)** direct separability. Values are AUROC (%).

because the FFN update is a direct sublayer update written to the residual stream and provides a clear default update source for analysis.

5.3 Direct Separability of GRA Trajectories

To examine whether GRA trajectories carry hallucination-relevant information before probe training, we evaluate the direct separability of their layer-wise statistics. Starting from the token-layer grounding trajectory $\mathbf{G} \in [0,1]^{m \times L}$, we first pool over generated-token positions using either Mean or Trend pooling, yielding one scalar statistic per layer. We then compute AUROC for each layer separately and report the best single-layer result.

As shown in Table 4**(c)**, single-layer trajectory statistics already achieve non-trivial separability on both representative models. This indicates that GRA trajectories contain hallucination-relevant information even before supervised probe training. Full GRA further improves the average performance by using the lightweight probe to aggregate information across layers and statistics, rather than relying on a single best layer.

6 Conclusion

In this paper, we introduced GRA, a white-box hallucination detection framework based on prompt-anchored residual update trajectories. GRA measures whether layer-wise internal updates remain anchored to the prompt sequence or shift toward the generated prefix during decoding. With mean-trend pooling and a lightweight probe, GRA produces a response-level reliability score for generated answers.

Beyond improving detection performance across multiple LLMs and datasets, GRA offers an interpretable view of hallucination as a change in internal update anchoring during generation. Our analyses show that this signal emerges over the token-level trajectory, appears across different residual update sources, and is already partially separable before supervised probe training. These findings suggest that prompt-anchored update dynamics provide an effective perspective for white-box hallucination detection. GRA is not a direct factuality verifier. Future work may extend this update-anchoring perspective to reasoning-oriented models and finer-grained token-level reliability estimation.

AI USE STATEMENT

We used generative AI tools to assist with manuscript preparation, including language editing and LaTeX formatting. All AI-assisted text, analyses, and claims were reviewed and verified by the authors, who take full responsibility for the final content of this submission.

REPRODUCIBILITY STATEMENT

The supplementary material specifies the models, datasets, data splits, seeds, hardware, evaluation protocols, probe architecture, training configuration, baseline details, prompt templates, complete ablations, and robustness analyses. Anonymous code and dependencies are included with the supplementary material.

REFERENCES

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. *arXiv preprint arXiv:2212.03827*, 2022.

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.

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

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

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, 2019.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

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, 2023.

Shize Liang and Hongzhi Wang. Neural probe-based hallucination detection for large language models. *arXiv preprint arXiv:2512.20949*, 2025.

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, 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.

Wang Ma, Debarun Bhattacharjya, Junkyu Lee, Nhan H Pham, Harsha Kokel, and Qiang Ji. Black-box uncertainty quantification for large language models via ensemble-of-ensembles. In *AAAI 2026 Workshop on Assessing and Improving Reliability of Foundation Models in the Real World*, 2026.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.

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.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Oscar Obeso, Andy Arditi, Javier Ferrando, Joshua Freeman, Cameron Holmes, and Neel Nanda. Real-time detection of hallucinated entities in long-form generation. *arXiv preprint arXiv:2509.03531*, 2025.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018.

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. *arXiv preprint arXiv:2209.15558*, 2022.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. A survey of uncertainty estimation methods on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21381–21396, 2025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuan-Jing Huang. Do large language models know what they don't know? In *Findings of the association for Computational Linguistics: ACL 2023*, pp. 8653–8665, 2023.

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 (Volume 1: Long Papers)*, pp. 17986–18002, 2025.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2):1–124, 2023.