WHAT MATTERS, NOT WHEN: DECIPHERING PROMPT REUSE WITH DISTRIBUTIONAL PROFILING FOR HALLUCINATION DETECTION IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Although large language models (LLMs) have demonstrated remarkable reasoning capabilities across diverse domains, they remain prone to hallucinations that undermine their reliability. To detect hallucinations, existing approaches generally leverage hidden states or attention matrices produced during generation as discriminative signals. Despite their effectiveness, these methods often implicitly assume that hallucination evidence is primarily encoded in generated tokens and their sequential organization. However, our empirical analysis reveals that hallucination risk can already be predicted before decoding, while detection remains largely insensitive to explicit temporal ordering. Motivated by these observations, we propose a novel framework, termed Distributional Profiling of Prompt Reuse (DOOR), for hallucination detection in LLMs. The core idea of DOOR is to characterize prompt reuse through the attention between prompt and generated tokens along two complementary dimensions: *reuse intensity* and *reuse dispersion*. More specifically, we first construct a distributional profile of prompt reuse from the attention received by prompt tokens during generation. Subsequently, we leverage mean reception and coefficient of variation to quantify reuse intensity and reuse dispersion, respectively, abstracting away from the temporal order of generation. To accommodate prompts of varying lengths, DOOR further introduces a rank-based profiling module that yields fixed-dimensional features, which are subsequently fed into a logistic regression probe for hallucination detection. Extensive experiments across diverse model backbones and benchmarks demonstrate the effectiveness and generalizability of DOOR. The code is available at https://anonymous.4open.science/r/DOOR-BE31/.

1 INTRODUCTION

Large language models (LLMs) have demonstrated strong performance across a wide range of natural language tasks (OpenAI, 2023; Team, 2025) and are increasingly being deployed in settings where reliability matters. Despite the impressive capabilities and broad utility of large language models (LLMs), their tendency to generate hallucinations remains a significant challenge to safe and trustworthy deployment (Ji et al., 2023; Huang et al., 2025; Zhang et al., 2023; Manakul et al., 2023; Kossen et al., 2024). Unlike errors made by traditional task-specific models, LLM hallucinations often manifest as plausible but factually incorrect, unsupported, or nonsensical content, making them difficult to identify solely from the generated text (Orgad et al., 2025).

A growing body of work addresses this problem by detecting hallucinations from internal model signals, such as hidden states (Bar-Shalom et al., 2025; Orgad et al., 2025; Zhang et al., 2025; Chatterjee et al., 2025; Azaria & Mitchell, 2023) and attention patterns (Chuang et al., 2024; Qi et al.; van Dijk, 2026). These methods are motivated by the observation that intrinsic representations are closely related to the extrinsic behavior of LLMs (Zhang et al., 2023; Han et al., 2025; Chen et al., 2024a). For example, existing approaches commonly train classifiers on hidden activations from different layers or characterize temporal variations in attention (Alansari et al., 2026; Srey et al., 2026; Qi et al.). Yet most existing studies focus on whether these signals are predictive, rather than on what structural properties of the signals carry the predictive information. However, it remains

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.

unclear which properties of these internal signals are fundamentally associated with hallucinations. Understanding this relationship could guide the development of more effective hallucination detectors.

In this work, we rethink what matters in representation-level hallucination detection. Some hallucination detectors assume that hallucination signals emerge and accumulate during autoregressive generation, following the so-called hallucination snowball effect (Zhang et al., 2024; Chen et al., 2024b; Li et al., 2026; Dou et al., 2024). Rather than taking this temporal view as given, we systematically examine when hallucination-discriminative information becomes accessible and how detection depends on the temporal order. Our analysis reveals two observations. First, *hallucination risk is already detectable before decoding*: prompt-side representations contain strong signals predictive of whether the subsequent response will hallucinate. Across generation stages, these signals evolve in a highly task-dependent manner rather than exhibiting a uniform accumulation pattern, and prompt-side internal states can be as informative as, or more informative than, generated-token representations. Second, *hallucination-discriminative signals persist after token-order destruction*: disrupting the native temporal order used by existing representation-level detectors does not systematically degrade their performance. Together, these findings reveal that hallucination signals are not confined to a progressively accumulating, explicitly ordered generation trajectory.

Motivated by these findings, we challenge the prevailing generation-centric view of hallucination detection (Su et al., 2024; Quevedo et al., 2024; Farquhar et al., 2024) and adopt a prompt-centric perspective. Our key insight is that the distribution of prompt reuse can retain hallucination signals even when eplicit generation order is discarded. We propose a simple but effective framework named Distributional Profiling of Prompt Reuse (DOOR), which characterizes how generated tokens reuse information from the prompt. Specifically, DOOR extracts generated-to-prompt attention and models prompt reuse through two descriptors: reuse intensity, which captures how strongly each prompt token is reused, and reuse dispersion, which captures how consistently that reuse varies across generation. It then independently ranks the two descriptors and samples them at fixed ranks, converting variable prompt statistics into a fixed-dimensional representation. The resulting profiles are concatenated across heads and layers and fed into a logistic-regression probe for hallucination detection. Across six benchmarks and three model backbones, DOOR achieves the strongest average AUROC and AUPRC among the evaluated competitive for each backbone. Cross-dataset transfer and within-question comparisons further show that its prompt-reuse profiles capture signals that extend beyond a single training distribution or prompt-level difficulty. We summarize our contributions as follows:

❶ ***New Analysis and Findings.*** We systematically investigate where hallucination-discriminative signals are accessible and how they depend on temporal ordering. We reveal that these signals are present before decoding and persist after order disruption.

❷ ***Novel Methodology and Theoretical Analysis.*** We propose DOOR, a prompt-centric framework that represents generated-to-prompt attention through the complementary distributions of reuse intensity and reuse dispersion. We theoretically characterize its permutation invariance, the complementary information captured by its two descriptors, and the approximation behavior of its profiles.

❸ ***Extensive Empirical Validation.*** We evaluate DOOR on diverse benchmarks and models. DOOR achieves the strongest average detection performance among competitive baselines. Cross-dataset transfer, length-controlled evaluation, and within-question comparisons further examine its generalization and its ability to distinguish responses beyond prompt-level difficulty.

2 Empirical Analysis: What Matters and When?

2.1 Problem Formulation

Following prior hallucination-detection literature (Orgad et al., 2025; Manakul et al., 2023; Li et al., 2023), we adopt a broad operational definition of hallucination that encompasses general response errors. Accordingly, responses judged to be incorrect are labeled as hallucinatory. Let $x = (x_1, \ldots, x_P)$ denote the tokenized prompt after applying the model's chat template, and let $o = (y_1, \ldots, y_T)$ denote the response generated by a frozen autoregressive language model. Each prompt–response pair is associated with a binary label $z \in \{0, 1\}$, where $z = 1$ indicates that the response is incorrect or hallucinatory and $z = 0$ indicates that it is faithful. Our goal is to learn a scoring function $S(x, o) \in [0, 1]$ that assigns higher scores to unreliable responses than to faithful ones. We consider a setting where the attention maps of the LLMs are accessible.

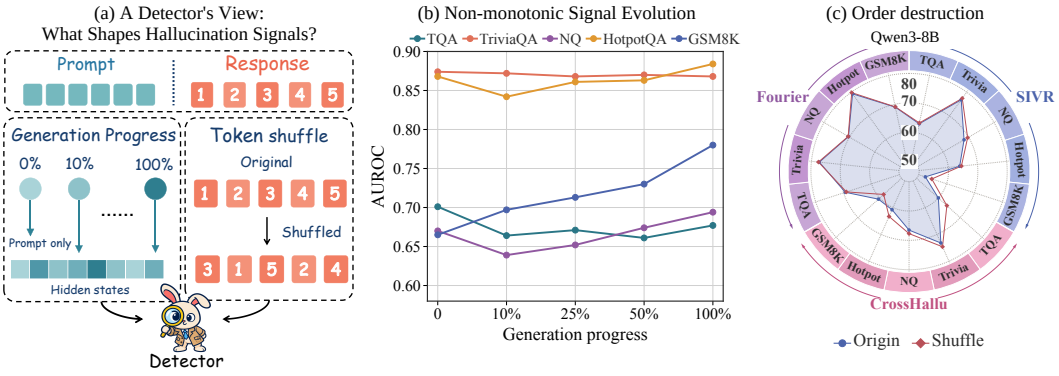


Figure 1: (a) Illustration of generation-stage probing and order disruption in detector inputs. (b) Hallucination-discriminative signals evolve non-monotonically during generation, as reflected by the AUROC of a linear probe across five benchmarks. (c) Disrupting the ordering used by each detector does not systematically degrade hallucination-detection performance.

2.2 Diagnostic Questions

Existing representation-level hallucination detectors commonly exploit generation-side dynamics, consistent with the hallucination-snowball view that errors can emerge and accumulate as autoregressive generation proceeds (Zhang et al., 2024; Su et al., 2024; Marfurt & Henderson, 2022; Jiang et al., 2024; Fadeeva et al., 2024). Nevertheless, hallucination in a large language model is also related to the interaction between the intrinsic knowledge of the model and the extrinsic prompt context (Longpre et al., 2021; Zhou et al., 2023; Chen et al., 2022; Tao et al., 2024; Xu et al., 2024). Therefore, we revisit the locus of hallucination and pose the following question: *Where are hallucination-discriminative signals accessible, and how do they evolve during generation?* For hallucination detection, what really matters and when?

Obs ❶:Hallucination Can Be Predicted Before Decoding. To examine how hallucination signals evolve during the generation process, we extract the hidden states from a middle layer for every token as features and train hallucination classifiers using logistic regression (linear probe) and an MLP (MLP probe) across datasets, models, and different fractions of generation progress. (Orgad et al., 2025; Azaria & Mitchell, 2023). At 0% progress, the probe reads only the prefilling hidden states of prompt tokens,therefore all accessible hallucination-discriminative signals are prompt-side. The remaining stages contain prompt-token representations together with incrementally generated tokens. Experimental details and full results are provided in Appendix D.2. As demonstrated in Figure 1 (b) and Figure 8, we observe that: *(i)* the hidden states available before the first generated token is decoded contain strong hallucination-discriminative signals. *(ii)* adding generated-token representations does not uniformly improve probe performance but instead follows strongly task-dependent trajectories. and *(iii)* generated tokens can still provide additional predictive information beyond the prompt, particularly for reasoning-intensive tasks such as GSM8K.

Obs ❷:Hallucination Signals Survive Token Order Destruction. To verify whether temporal information contributes to hallucination detection, we consider three baselines: Semantic Information Variation Rate (SIVR) (Srey et al., 2026), which assumes that variation rates across representation levels change over the course of generation. CrossHallu (Alansari et al., 2026), which estimates hallucination from the magnitudes of generation-centric token-probability transitions, and Frequency-Aware Attention (Fourier) (Qi et al.), which applies a Fourier transform to attention vectors along the attended-position axis. We then disrupt the ordering used by each detector and evaluate the resulting hallucination-detection performance. Experimental details and full results are provided in Appendix D.1. Figure 1 (c) and 7 show that, across representative baselines at different representation levels and from different perspectives, disrupting the native token order causes no systematic average degradation in hallucination-detection performance. Despite relying on different generation-side signals, the tested detectors retain comparable performance after order destruction. These results indicate that explicit temporal ordering plays only a limited role in the discriminative information exploited by these hallucination detectors.

Together, these observations challenge the view that hallucination detection must primarily rely on generation-centric temporal dynamics. These results suggest that the tested detectors can retain

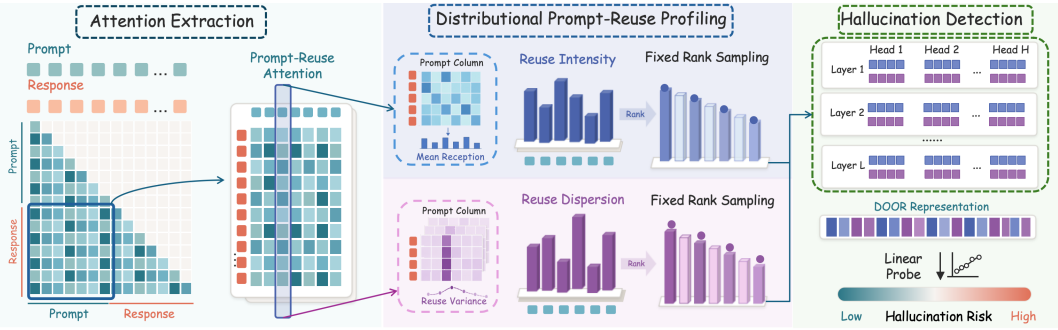


Figure 2: Overview of DOOR. DOOR extracts generated-to-prompt attention from each layer and head, summarizes prompt reuse using reuse intensity and reuse dispersion, and converts the variable-length descriptors into fixed-dimensional rank profiles through fixed-rank sampling. The profiles are then concatenated into the final representation and passed to a linear probe for hallucination detection.

hallucination-discriminative information after the specified order permutations, although the effects vary across methods and datasets. This shifts our attention from tracking a particular generation trajectory to tap into the distributional structure of prompt-response interaction. While the pre-filling analysis identifies predictive information on the prompt side, DOOR further characterizes how an observed response interacts with that prompt during generation. We therefore focus on prompt reuse, asking how prompt information is reused throughout generation.

3 The Proposed DOOR

3.1 Framework Overview

Motivated by the empirical observations above, DOOR shifts the focus from token-level generation trajectories to the distributional structure of prompt–response interaction. Rather than explicitly modeling generation order, DOOR characterizes how the response reuses information from the prompt. As Figure 2 shows, DOOR extracts the response-to-prompt attention block from a single forward pass and summarizes prompt reuse from two complementary perspectives: reuse intensity, which captures how strongly each prompt token is attended to, and reuse dispersion, which captures how unevenly attention is distributed across generation steps. Each profile is then ranked by value and sampled at fixed ranks. The sampled profiles from all attention heads and layers are combined into the DOOR representation and fed into a logistic regression probe for hallucination detection.

3.2 Distributional Profiling for Prompt Reuse Characterization

Prior work has shown that attention maps encode predictive signals for hallucination detection (Ogasa & Arase, 2025; Bazarova et al., 2026; Samaga et al., 2026). The key intuition is that the way generated tokens attend to the input prompt reflects how the model utilizes the provided information during generation. Therefore, we extract generated-to-prompt attention patterns as an intermediate representation for hallucination analysis. Consider layer $\ell \in \{1, \ldots, L\}$ and attention head $h \in \{1, \ldots, H\}$. Let $\mathbf{A}^{\ell,h} \in \mathbb{R}^{(P+T)\times(P+T)}$ denote the causal attention matrix of the prompt-response sequence (Vaswani et al., 2017). For each generated token y_t, we extract its attention to all prompt tokens as $\mathbf{p}_t^{\ell,h} = \mathbf{A}^{\ell,h}_{P+t,\,1:P} \in \mathbb{R}^P$, for $t = 1, \ldots, T$. Stacking these generated-to-prompt attention vectors gives $\mathbf{P}^{\ell,h} = [\mathbf{p}_1^{\ell,h}, \ldots, \mathbf{p}_T^{\ell,h}]^\top \in \mathbb{R}^{T\times P}$, where each row represents how a generated token attends to the prompt. Overall, $\mathbf{P}^{\ell,h}$ provides a view of prompt utilization during generation.

Mean Reception for Reuse Intensity. For each prompt position j, we compute the average attention it receives from the generated response:

$$\mu_j^{\ell,h} = \frac{1}{T}\sum_{t=1}^{T} p_{t,j}^{\ell,h}, \qquad \boldsymbol{\mu}^{\ell,h} = (\mu_1^{\ell,h}, \ldots, \mu_P^{\ell,h}). \tag{1}$$

The resulting vector $\boldsymbol{\mu}^{\ell,h}$ represents the column-wise mean of $\mathbf{P}^{\ell,h}$. A large value indicates that the corresponding prompt position receives greater attention on average across generated positions.

Crucially, we retain one statistic for each prompt position rather than collapsing the entire prompt into a single attention-mass scalar. This preserves the distribution of reuse across prompt positions and allows the representation to distinguish concentrated reuse from broadly distributed reuse.

Coefficient of Variation for Reuse Dispersion. Mean reception alone cannot distinguish persistent reuse from intermittent reuse. Two prompt positions may have identical mean reception while one is attended to consistently across generation and the other receives attention only at a small number of steps. We therefore compute the population standard deviation across generated tokens and normalize it by the corresponding mean reception:

$$\sigma_j^{\ell,h} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(p_{t,j}^{\ell,h} - \mu_j^{\ell,h}\right)^2}, \qquad c_j^{\ell,h} = \frac{\sigma_j^{\ell,h}}{\max\left(\mu_j^{\ell,h}, \varepsilon\right)}, \tag{2}$$

where $\varepsilon > 0$ is a numerical stabilizer for prompt positions with negligible mean reception. The resulting vector $\mathbf{c}^{\ell,h} = (c_1^{\ell,h}, \ldots, c_P^{\ell,h})$ provides a normalized measure of reuse dispersion. A large value indicates bursty reuse relative to the prompt token's own average reception.

3.3 Rank-Based Profiling for Fixed-Dimensional Representation

Because input length can systematically affect contextual evidence use and hallucination detection, prompt length may act as a nuisance factor (Liu et al., 2024; 2025; 2022). The prompt reuse descriptors have variable dimension P due to different prompt lengths. Padding introduces arbitrary empty coordinates, while global pooling discards the distributional structure. We therefore propose a rank-based profiling operator Π_K that maps a variable-length descriptor into a fixed-dimensional representation while preserving its distributional shape.

For a descriptor $\mathbf{u} \in \mathbb{R}^P$, let $u_{(1)} \geq u_{(2)} \geq \cdots \geq u_{(P)}$ denote its descending order statistics. We select K evenly spaced ranks and obtain the profile values through linear interpolation:

$$r_k = 1 + \frac{(k-1)(P-1)}{K-1}, \qquad [\Pi_K(\mathbf{u})]_k = (1-\alpha_k)u_{(a_k)} + \alpha_k u_{(\min\{a_k+1,P\})}, \tag{3}$$

where $a_k = \lfloor r_k \rfloor$ and $\alpha_k = r_k - a_k$. We apply Π_K independently to reuse intensity and reuse dispersion and concatenate the resulting profiles across all layers and heads:

$$\mathbf{x}_{\mathrm{RD+RI}} = \mathrm{vec}\left(\left\{[\Pi_K(\mathbf{c}^{\ell,h}); \Pi_K(\boldsymbol{\mu}^{\ell,h})]\right\}_{\ell,h}\right) \in \mathbb{R}^{2KLH}. \tag{4}$$

This yields a fixed-dimensional representation that retains the distributional structure of prompt reuse while discarding absolute token positions.

3.4 Linear Probing for Hallucination Detection

Given the extracted representation $\mathbf{x}_i$ and binary label z_i, we train a logistic regression probe to produce a hallucination score:

$$\hat{s}_i = \mathrm{sigmoid}\left(\mathbf{w}^\top \tilde{\mathbf{x}}_i + b\right). \tag{5}$$

where each feature dimension is standardized using statistics computed from the training split only. The probe output $\hat{s}_i$ is used directly as the hallucination score. To account for class imbalance, we use a class-balanced logistic regression with ℓ_2 regularization. The classifier is trained separately for each backbone model, while the language model and the proposed attention representation remain completely frozen. Therefore, the probe only serves as a mapping from the fixed `DOOR` representation to a hallucination score. The complete procedure is provided in Algorithm 1 in Appendix B.1.

3.5 Theoretical Analysis

We justify `DOOR` through three properties of its prompt-reuse representation: permutation invariance and scale separation, complementary identification by RI and RD, and finite approximation by rank profiling. These properties correspond to the order-destruction, component-ablation, and profile-length experiments, respectively. All proofs are deferred to Appendix E.

Proposition 3.1. *Let $\mathbf{P} \in \mathbb{R}_{\geq 0}^{T \times P}$ be a generated-to-prompt attention block, and let $\mathbf{x}_{\mathrm{RI}}(\mathbf{P})$ and $\mathbf{x}_{\mathrm{RD}}(\mathbf{P})$ be its independently ranked RI and RD profiles. For arbitrary permutation matrices $\mathbf{Q}_T$ and $\mathbf{Q}_P$, the joint representation $\mathbf{x}_{\mathrm{RD+RI}} := [\mathbf{x}_{\mathrm{RD}}; \mathbf{x}_{\mathrm{RI}}]$ satisfies*

$$\mathbf{x}_{\mathrm{RD+RI}}\left(\mathbf{Q}_T \mathbf{P} \mathbf{Q}_P^{\top}\right) = \mathbf{x}_{\mathrm{RD+RI}}(\mathbf{P}). \tag{6}$$

Remark 3.1. *Proposition 3.1 shows that* DOOR *records which reuse levels occur rather than their positions. By construction, RI retains absolute reception magnitude, whereas RD captures relative dispersion. Independent ranking removes the positional correspondence between the two profiles.*

Proposition 3.2. *Suppose a prompt position follows $p_{t,j} = a_j B_{t,j}$, where $a_j > 0$ is its active attention strength and $B_{t,j} \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(\rho_j)$ indicates reuse with frequency $0 < \rho_j \leq 1$. RI alone confounds a_j and ρ_j, whereas RD omits a_j; the population mean $\mu_j^\star$ and the unstabilized population coefficient of variation $c_j^\star$ jointly recover both:*

$$\mu_j^\star = a_j \rho_j, \qquad c_j^\star = \sqrt{\frac{1-\rho_j}{\rho_j}}. \tag{7}$$

where $\rho_j = 1/\big(1 + (c_j^\star)^2\big)$ and $a_j = \mu_j^\star\big(1 + (c_j^\star)^2\big)$.

Remark 3.2. *Proposition 3.2 shows that RI and RD are complementary at the per-token level under the marginal model, motivating their joint use in* DOOR*. Whether their token-wise correspondence should be preserved after rank profiling is evaluated empirically in Table 5.*

Theorem 3.3. *Let q be the full interpolated rank curve of a prompt-reuse descriptor, and let $\widehat{q}_K$ be its reconstruction from K uniformly sampled ranks. For a fixed prompt length, two descriptors have the same full rank curve if and only if their values differ only by a permutation. If q is L-Lipschitz and a distributional score S is γ-Lipschitz with respect to the supremum norm, then*

$$|S(q) - S(\widehat{q}_K)| \leq \gamma \left\| q - \widehat{q}_K \right\|_\infty \leq \frac{\gamma L}{2(K-1)}. \tag{8}$$

Remark 3.3. *Theorem 3.3 shows that ranking removes token positions while preserving the descriptor values. Moreover, the effect of sampling K ranks on any Lipschitz distributional score decreases as $O(K^{-1})$. Together, these results characterize the invariance, complementarity, and approximation properties underlying distributional representation in* DOOR*.*

4 Experiments

4.1 Experimental Settings

Benchmark Datasets. We evaluate hallucination detection on six widely used question-answering benchmarks covering factual knowledge, open-domain retrieval, and reasoning tasks: TruthfulQA (Lin et al., 2022), TriviaQA (Joshi et al., 2017), Natural Questions (NQ-Open) (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), GSM8K (Cobbe et al., 2021), Halueval (Li et al., 2023). These datasets represent diverse hallucination patterns, including factual inconsistency, knowledge uncertainty, multi-hop reasoning errors, and mathematical reasoning failures.

Compared Baselines. We compare DOOR with representative hallucination detection methods based on different internal and attention signals. These baselines include attention-based approaches such as Lookback Lens (Chuang et al., 2024), AttentionDvg (van Dijk, 2026), FourierAttention (Qi et al.), and LapEigVals (Binkowski et al., 2025), as well as hidden-state probing methods including ICR (Zhang et al., 2025), LLMs Know (Orgad et al., 2025), ACT-ViT (Bar-Shalom et al., 2025), and HaMI (Niu et al., 2025). The baseline details are provided in Appendix A.2.

Evaluation and Implementation Details. We evaluate DOOR on three instruction-tuned model families: Qwen3-8B (Yang et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024), and Mistral-7B (Jiang et al., 2023). Responses are generated using the same decoding protocol, and their attention representations are extracted through a teacher-forced forward pass. We extract generated-to-prompt attention from all transformer layers and convert the RI and RD descriptors into fixed-length rank profiles with $K = 8$ and $\epsilon = 10^{-8}$. Following prior hallucination detection studies (Orgad et al., 2025; Chuang et al., 2024), we report AUROC and AUPRC, treating incorrect responses as the positive class. Results are reported as mean and standard deviation over three question-disjoint split seeds. Further protocol and implementation details are provided in Appendix A.1.

Table 1: Main hallucination-detection results across six benchmarks and three LLM backbones. We report mean AUROC/AUPRC (%) ± standard deviation. Dark and light blue cells indicate the best and second-best results, respectively, within each backbone–dataset setting.

Backbone	Method	TruthfulQA		TriviaQA		NQ-Open		HotpotQA		GSM8K		HaluEval		Avg.	
		AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Qwen3-8B	Lookback Lens	68.1 ±2.6	90.5 ±0.3	80.6 ±1.6	73.4 ±5.4	70.3 ±0.3	76.7 ±1.7	82.9 ±1.5	90.8 ±1.7	75.1 ±4.5	88.0 ±2.9	76.4 ±2.4	82.2 ±3.0	75.6 ±1.0	83.6 ±1.2
	AttentionDvg	68.5 ±2.7	91.0 ±0.5	83.6 ±1.1	75.3 ±3.0	72.6 ±0.6	79.4 ±0.9	82.1 ±0.3	89.4 ±2.6	72.6 ±3.3	87.0 ±2.2	76.3 ±1.8	83.0 ±2.2	76.0 ±0.8	84.2 ±0.8
	FourierAttention	71.1 ±1.0	91.5 ±0.8	81.3 ±0.5	73.0 ±4.7	73.6 ±0.8	78.8 ±1.0	80.0 ±1.1	88.4 ±2.0	76.0 ±4.4	89.1 ±2.3	76.0 ±1.9	82.9 ±1.4	76.3 ±0.8	84.0 ±0.9
	LapEigVals	69.7 ±0.2	91.6 ±0.6	83.5 ±2.4	78.1 ±4.4	72.3 ±2.1	79.9 ±0.7	78.2 ±4.2	86.1 ±2.0	75.3 ±2.8	89.0 ±1.2	71.3 ±0.7	79.1 ±2.8	75.0 ±1.0	83.9 ±1.0
	ICR	57.9 ±3.2	86.7 ±2.5	74.6 ±3.6	65.4 ±5.1	66.5 ±0.9	75.3 ±2.3	60.0 ±5.9	76.0 ±1.8	68.2 ±4.0	82.5 ±4.0	59.8 ±0.4	69.3 ±3.1	64.5 ±1.4	75.9 ±1.4
	LLMs Know	68.7 ±3.4	91.3 ±1.9	79.0 ±2.1	71.9 ±5.3	71.7 ±2.7	76.8 ±4.6	80.7 ±2.4	88.1 ±4.2	67.5 ±3.7	83.6 ±2.9	77.4 ±1.5	82.9 ±0.6	74.1 ±1.1	82.4 ±1.5
	ACT-ViT	69.5 ±3.5	91.7 ±1.0	83.9 ±1.4	78.0 ±1.9	71.3 ±2.8	79.5 ±3.8	81.2 ±3.5	88.1 ±4.7	61.4 ±6.1	80.4 ±3.5	78.6 ±0.6	82.4 ±1.4	74.4 ±1.4	83.3 ±1.3
	HaMI	51.0 ±2.8	83.9 ±1.5	73.1 ±3.6	66.6 ±4.2	71.2 ±1.3	78.2 ±2.0	76.5 ±1.9	85.2 ±2.7	64.6 ±6.5	83.2 ±4.3	70.0 ±8.2	75.8 ±6.6	67.8 ±2.0	78.8 ±1.6
	DOOR	**80.8** ±2.0	**95.0** ±0.9	**85.6** ±0.6	**79.0** ±1.3	**75.2** ±1.5	**80.1** ±2.5	**85.8** ±1.5	**92.3** ±3.1	**79.1** ±5.0	**89.7** ±3.5	**79.7** ±2.3	**84.9** ±3.4	**81.0** ±1.1	**86.8** ±1.1
Llama-3.1-8B	Lookback Lens	71.7 ±7.5	89.3 ±4.0	88.6 ±1.5	81.8 ±3.4	69.6 ±0.5	63.5 ±3.0	86.1 ±1.2	96.3 ±0.7	79.1 ±2.9	58.7 ±9.8	83.3 ±2.1	90.1 ±2.7	79.7 ±1.5	79.9 ±2.0
	AttentionDvg	71.8 ±4.6	89.8 ±2.6	89.2 ±1.8	81.8 ±2.9	70.6 ±2.1	65.5 ±3.6	87.3 ±2.9	96.3 ±1.5	77.3 ±2.5	50.3 ±6.9	83.8 ±0.7	90.6 ±1.0	80.0 ±1.1	79.0 ±1.5
	FourierAttention	70.1 ±5.3	88.9 ±2.4	89.3 ±2.9	82.5 ±4.7	71.1 ±1.2	64.5 ±1.7	86.1 ±0.8	96.2 ±0.5	82.7 ±2.1	56.2 ±7.4	85.7 ±1.1	92.2 ±0.3	80.9 ±1.1	80.1 ±1.6
	LapEigVals	70.5 ±4.1	90.9 ±1.7	90.1 ±1.4	83.0 ±5.3	67.5 ±4.5	63.8 ±6.5	86.9 ±2.6	96.4 ±0.7	80.5 ±1.2	57.7 ±1.1	85.6 ±1.3	92.5 ±1.4	80.2 ±1.2	80.8 ±1.4
	ICR	63.2 ±4.0	86.9 ±3.2	81.2 ±3.7	72.2 ±6.1	64.6 ±1.3	59.3 ±1.9	82.2 ±2.4	95.5 ±0.4	67.2 ±2.4	46.4 ±5.7	79.8 ±2.7	88.8 ±2.5	73.0 ±1.2	74.9 ±1.6
	LLMs Know	65.9 ±0.8	87.7 ±0.4	84.3 ±5.7	74.1 ±11.5	**71.3** ±3.2	66.3 ±3.4	88.1 ±5.1	95.5 ±3.4	83.1 ±2.8	66.1 ±4.1	87.5 ±0.3	92.7 ±0.2	80.0 ±1.5	80.4 ±2.2
	ACT-ViT	67.9 ±0.9	89.0 ±1.7	88.2 ±1.7	83.2 ±3.9	66.7 ±5.6	63.8 ±5.1	89.2 ±3.3	97.5 ±0.7	67.4 ±1.3	40.3 ±5.5	86.1 ±1.6	92.4 ±1.8	77.6 ±1.2	77.7 ±1.4
	HaMI	65.8 ±7.5	87.1 ±4.2	82.6 ±4.7	79.0 ±4.7	64.2 ±2.6	64.3 ±2.7	88.6 ±1.2	97.2 ±0.4	74.5 ±7.0	51.1 ±9.4	83.7 ±2.7	90.9 ±2.6	76.5 ±2.0	78.2 ±2.0
	DOOR	**76.7** ±4.7	**92.4** ±1.5	**90.4** ±2.3	**85.2** ±4.1	69.9 ±0.7	**68.3** ±0.4	**90.5** ±2.2	**97.8** ±0.4	**84.2** ±3.3	**66.3** ±5.3	**88.0** ±1.8	**94.1** ±1.6	**83.2** ±1.1	**84.0** ±1.2
Mistral-7B	Lookback Lens	73.2 ±3.9	92.6 ±0.7	76.6 ±1.4	62.8 ±1.8	67.1 ±1.1	68.9 ±2.3	70.0 ±2.1	80.3 ±4.7	78.2 ±1.7	83.9 ±3.7	72.3 ±2.2	76.8 ±3.1	72.9 ±0.9	77.5 ±1.2
	AttentionDvg	70.0 ±3.0	91.9 ±2.4	75.9 ±1.2	60.1 ±0.7	65.5 ±2.5	70.0 ±3.2	71.4 ±1.5	79.7 ±2.6	79.4 ±0.7	84.2 ±2.1	71.7 ±5.0	77.2 ±6.6	72.3 ±1.1	77.2 ±1.4
	FourierAttention	69.1 ±5.6	91.2 ±2.1	76.6 ±0.9	61.5 ±3.2	**69.5** ±1.7	**73.0** ±0.6	70.5 ±2.4	81.0 ±1.5	80.7 ±2.6	84.2 ±3.7	74.1 ±3.1	78.3 ±3.5	73.4 ±1.3	78.2 ±1.1
	LapEigVals	67.9 ±5.4	91.4 ±0.6	75.5 ±1.2	58.1 ±3.8	67.6 ±1.5	70.0 ±0.9	69.7 ±1.5	82.3 ±2.4	81.6 ±1.8	84.8 ±0.9	74.8 ±2.6	79.6 ±2.1	72.9 ±1.1	77.7 ±0.8
	ICR	67.5 ±7.6	91.1 ±1.1	62.1 ±4.7	43.0 ±4.8	60.3 ±4.9	61.8 ±6.8	59.4 ±3.0	74.8 ±1.3	58.6 ±2.3	67.3 ±4.8	64.6 ±1.4	70.5 ±4.5	62.1 ±1.8	68.1 ±1.8
	LLMs Know	74.8 ±7.1	93.5 ±1.2	77.0 ±1.3	59.1 ±0.6	63.9 ±5.7	68.4 ±6.3	74.4 ±1.0	84.5 ±1.8	72.4 ±0.5	77.2 ±2.8	72.1 ±5.3	76.6 ±6.5	72.4 ±1.8	76.5 ±1.7
	ACT-ViT	71.7 ±3.0	92.4 ±1.2	77.7 ±4.6	64.6 ±7.0	61.4 ±2.2	65.8 ±2.4	74.7 ±1.8	**84.9** ±3.4	74.6 ±2.1	79.2 ±3.1	75.6 ±0.6	80.8 ±2.6	72.6 ±1.1	78.0 ±1.6
	HaMI	61.0 ±5.9	88.0 ±4.7	75.4 ±7.8	63.9 ±8.2	63.4 ±5.0	66.4 ±7.1	72.9 ±0.8	81.5 ±4.0	76.4 ±2.8	82.0 ±4.3	74.8 ±2.2	80.8 ±1.3	70.6 ±2.0	77.1 ±2.2
	DOOR	**78.8** ±4.9	**94.2** ±0.4	**78.6** ±0.6	**66.0** ±1.0	**69.5** ±1.3	72.0 ±1.8	**75.0** ±1.8	83.0 ±0.5	**81.7** ±1.2	**86.1** ±1.5	**80.3** ±3.2	**84.2** ±4.6	**77.3** ±1.1	**81.0** ±0.9

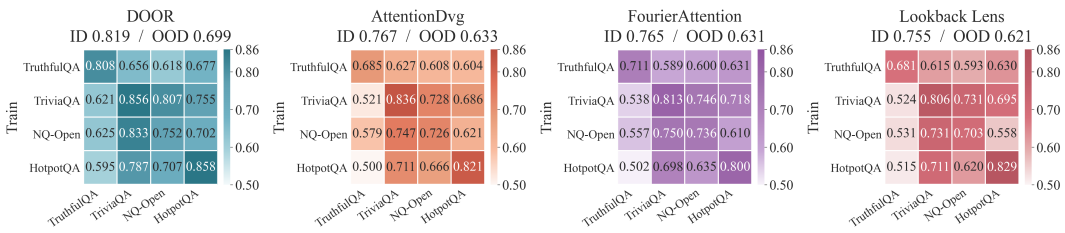


Figure 3: Cross-dataset generalization of hallucination detection methods. Each heatmap reports AUROC when training on one dataset and evaluating on another dataset. Results are shown for DOOR and representative attention-based baselines across four benchmarks.

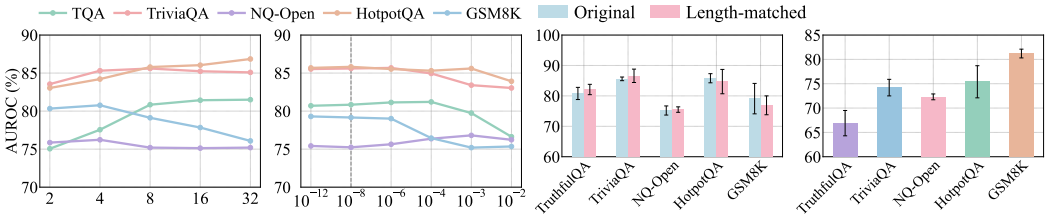
4.2 Empirical Results

In-Distribution Performance. To evaluate the effectiveness of prompt-reuse signals under matched training and test distributions, we compare DOOR with representative hidden-state and attention-based detectors across six benchmarks and three backbones. Three observations emerge from Table 1. *First,* DOOR *achieves the strongest average performance across all three backbones*, demonstrating that its effectiveness extends across the evaluated model families. This consistency suggests that the distributional patterns captured by DOOR are not specific to the attention behavior of a single backbone. *Second, distributional prompt-reuse statistics are competitive across both factual and reasoning benchmarks*. The gains are therefore not confined to a single dataset or error type, despite differences in the knowledge and reasoning demands of these tasks. *Third, attention-based signals are generally more reliable than the hidden-state probes considered here*, although the relative ordering varies by backbone and dataset. Taken together, these results establish prompt-reuse profiling as an effective way to extract discriminative information from model attention.

Out-of-Distribution Performance. To examine whether the detected signals generalize beyond the training distribution, we compare four attention-based methods when trained and evaluated on different factual QA benchmarks using Qwen3-8B. Two observations emerge from Figure 3. *First,* DOOR *exhibits stronger cross-dataset transfer than the compared attention-based methods*. Its advantage across multiple transfer directions suggests that the distributional structure of prompt reuse captures information shared across different question distributions, rather than only patterns specific to the training benchmark. *Second, hallucination detection remains affected by dataset dependency*: average transfer performance is lower than performance under matched training and test distributions for all compared methods. Thus, DOOR improves the transferability of the detected signal while leaving substantial room for further progress under domain shift.

Table 2: Comparison of controlled variants of DOOR on Qwen3-8B. We report AUROC and AUPRC (%) as mean ± standard deviation over three question-disjoint split seeds. Dark and light blue cells indicate the best and second-best results, respectively, within each dataset setting.

	TruthfulQA		TriviaQA		NQ-Open		HotpotQA		GSM8K		Avg.	
Variant	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
V1	79.8 ±2.8	94.9 ±1.0	84.6 ±1.0	77.9 ±1.5	73.9 ±2.6	79.0 ±3.4	84.8 ±0.7	92.1 ±2.6	77.9 ±3.3	89.1 ±3.6	80.2 ±1.0	86.6 ±1.2
V2	73.2 ±2.2	91.7 ±2.6	83.6 ±0.9	76.6 ±2.6	73.0 ±1.2	79.9 ±1.7	83.2 ±2.2	90.4 ±3.3	76.7 ±4.5	88.7 ±1.7	77.9 ±1.1	85.5 ±1.1
V3	76.8 ±2.1	93.6 ±1.2	83.3 ±1.8	75.8 ±2.8	**75.2** ±0.6	77.5 ±2.0	83.4 ±1.9	90.7 ±3.4	76.9 ±4.9	88.8 ±2.0	79.1 ±1.2	85.3 ±1.1
V4	75.1 ±3.1	92.8 ±1.2	79.4 ±2.7	76.8 ±1.5	69.3 ±2.2	77.5 ±1.8	79.0 ±3.4	90.1 ±2.0	73.2 ±2.9	86.9 ±1.6	75.2 ±2.5	84.8 ±1.4
V5	78.4 ±2.2	94.1 ±1.3	84.8 ±2.1	78.3 ±3.7	74.8 ±2.1	**80.3** ±2.6	85.1 ±2.0	92.0 ±3.8	76.8 ±4.1	87.4 ±2.8	80.0 ±1.2	86.4 ±1.3
V6	72.0 ±3.0	92.8 ±0.8	84.8 ±3.3	77.3 ±7.4	73.2 ±0.8	79.3 ±0.9	80.1 ±0.8	87.4 ±1.3	76.2 ±2.0	89.2 ±1.5	77.3 ±1.0	85.2 ±1.6
DOOR	**80.8** ±2.0	**95.0** ±0.9	**85.6** ±0.6	**79.0** ±1.3	**75.2** ±1.5	80.1 ±2.5	**85.8** ±1.5	**92.3** ±3.1	**79.1** ±5.0	**89.7** ±3.5	**81.3** ±1.2	**87.2** ±1.1



(a) Effect of Profile Length K (b) Effect of Numerical Stabilizer ε (c) Length-Matched Evaluation (d) Beyond-Prompt Hallucination

Figure 4: Further analysis of DOOR. (a) Sensitivity to profile length K. (b) Sensitivity to the numerical stabilizer ϵ. (c) Length-controlled evaluation by matching hallucinated and faithful responses with similar prompt and response lengths. (d) Beyond-prompt hallucination detection using multiple sampled responses from the same prompt to evaluate response-conditioned signals.

Ablation Studies. We compare DOOR with six controlled variants. V1 retains only the reuse-dispersion (RD) channel, while V2 retains only the reuse-intensity (RI) channel. V3 replaces the coefficient of variation used in RD with the raw standard deviation. V4 replaces each RI and RD rank profile with its mean and standard deviation. V5 preserves the token-wise correspondence between RI and RD by sorting prompt positions according to RI and sampling RD at the same RI-induced ranks. V6 computes both profiles from generated-to-generated attention instead of generated-to-prompt attention. The full results are in Table 5. *First, RD provides the stronger individual channel, while RI remains complementary*: the full representation attains the highest average AUROC for all backbones. *Second, replacing each rank profile with its mean and standard deviation reduces performance.* This result is consistent with the rank profiles retaining useful predictive information. *Third, generated-to-prompt attention is the more reliable source*: moving the same descriptors to the generation to generation block reduces performance.

Sensitivity Analysis. To evaluate the robustness of DOOR, we conduct sensitivity analyses with respect to the profile length K in the rank-based representation and the numerical stabilizer ϵ, as shown in Figure 4. *First, there is a tradeoff between profile resolution and representation complexity.* Increasing K improves the average performance up to a moderate resolution, while further increasing the profile length provides limited additional gains. The effect also varies across datasets: NQ-Open remains relatively stable, whereas GSM8K shows a gradual performance decrease with larger K. These results suggest that overly short profiles may lose useful distributional structures, while excessive resolution does not necessarily improve discrimination. *Second, the numerical stabilizer has limited influence on performance.* The detection results remain stable across a broad range of small ϵ values, indicating that DOOR is not sensitive to the choice of numerical stabilization.

Length-Controlled Evaluation. To examine whether the effectiveness of DOOR can be attributed to prompt or response length differences, we conduct a length-controlled evaluation by matching hallucinated and faithful responses with similar prompt and response lengths. The full details are in Appendix C.1. As illustrated in Figure 4 (c), it is noticeable that DOOR *remains robust after controlling for length-related factors.* The performance remains comparable between the original and length-matched settings across different benchmarks, suggesting that the detected hallucination signals are not primarily explained by superficial length statistics.

Beyond-Prompt Hallucination Detection. To investigate whether DOOR captures response-specific signals beyond prompt-level information, we perform a multiple-generation evaluation with fixed prompts using within-question AUROC. For each question, multiple responses are sampled, where they share the same prompt but may differ in factual correctness. The full details are in Appendix C.2. Figure 4(d) shows that DOOR *captures response-conditioned hallucination signals beyond prompt-level*

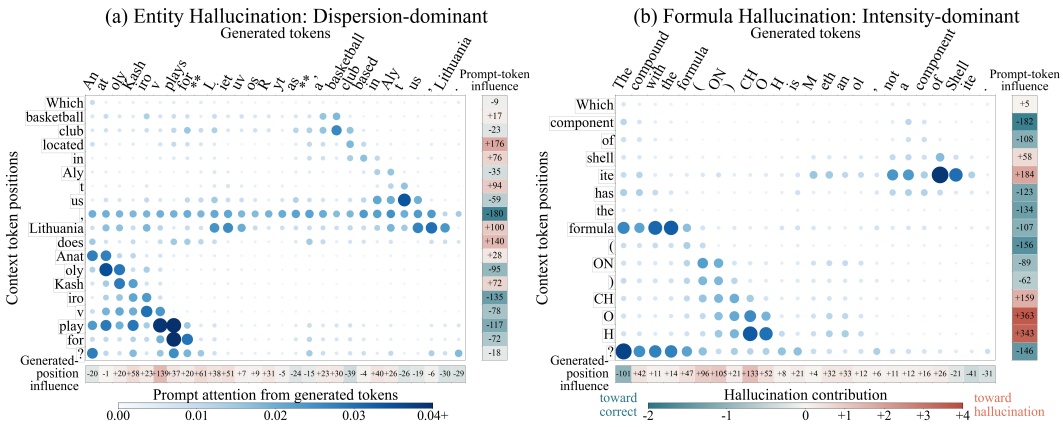


Figure 5: Case studies captured by DOOR. Each case-study panel visualizes the generated-to-prompt attention block, where the horizontal axis denotes generated tokens and the vertical one denotes prompt. The strips show the corresponding prompt-token and generated-position attribution scores. For readability, the attribution values shown in the side strips are multiplied by 100.

uncertainty. Since prompt information remains constant within each group, successful response-level discrimination requires additional evidence from the generated response.

Case Analysis. To illustrate the evidence captured by DOOR, we visualize two representative hallucination cases with different dominant reuse patterns: one RD-dominant case and one RI-dominant case. Figure 5 shows the generated-to-prompt attention blocks together with the attribution scores. The visualization details are in Appendix D.3. Two observations emerge: *First, hallucinated responses can still exhibit substantial prompt utilization*. Strong generated-to-prompt attention does not necessarily indicate correctness. *Second, hallucination cases exhibit distinct reuse profiles*, with some responses showing dispersed reuse patterns and others relying on concentrated prompt regions.

5 Related Work

Hallucinations in LLMs. LLM hallucinations include plausible but factually incorrect content (Zhang et al., 2023), claims unsupported by the input context (Niu et al., 2024; Pan et al., 2026; Maynez et al., 2020), or statements inconsistent with external knowledge. They remain a challenge to reliable language systems (Ji et al., 2023; Huang et al., 2025). Prior work studies their taxonomy, causes, and mitigation strategies (Wang et al., 2024). Such errors can arise from gaps in learned knowledge, insufficient grounding, and reasoning failures (Zhang et al., 2023).

Hallucination Detection via Internal Signals. Internal LLM signals can provide indicators of hallucination (Azaria & Mitchell, 2023; Chatterjee et al., 2025). For example, Chuang et al. (2024) measure attention allocation between generated and prompt tokens, while van Dijk (2026) study divergence in attention distributions. Other work characterizes attention matrices through Fourier-domain features or spectral properties of attention graphs (Qi et al.; Binkowski et al., 2025). Hidden representations offer a complementary view: Azaria & Mitchell (2023) show that hidden states encode reliability information, while others examine layer-wise dynamics, intrinsic representations, activation trajectories, and adaptive token selection (Zhang et al., 2025; Orgad et al., 2025; Bar-Shalom et al., 2025; Niu et al., 2025). DOOR extends attention-based methods by profiling the distribution of attention received by prompt tokens during generation.

6 Conclusion

In this work, we revisit what matters in representation-level hallucination detection. Our analysis reveals that hallucination can already be predicted before decoding and that shuffling the tested token orderings does not systematically reduce detection performance. Based on these observation, we propose DOOR, which characterizes prompt reuse through complementary prompt reuse descriptors and represents them with fixed-dimensional rank profiles. Extensive experiments across multiple models and benchmarks demonstrate that DOOR provides effective hallucination detection and generalization.

AI USE STATEMENT

Generative AI tools were used solely for language polishing and the correction of grammatical and stylistic errors in the manuscript. They were not used to formulate the research questions, develop the proposed method, design or conduct the experiments, analyze the results, or draw the scientific conclusions. The authors reviewed and verified all AI-assisted edits and take full responsibility for the content of the paper.

ETHICS STATEMENT

This work studies the detection of hallucinations in large language models with the goal of improving their reliability. All experiments are conducted using publicly available language models and established research benchmarks; no private, personally identifiable, or sensitive user data are collected. Nevertheless, hallucination detectors are imperfect and should not be treated as guarantees of factual correctness. In safety-critical applications, their predictions should be combined with external verification and appropriate human oversight. We report aggregate experimental results and discuss the limitations of the proposed detector to reduce the risk of misleading interpretations or inappropriate downstream use.

REPRODUCIBILITY STATEMENT

We provide code required to reproduce the experiments and results reported in this paper. The submission includes the complete implementation of `DOOR`, together with the corresponding training, evaluation.

REFERENCES

Aisha Alansari, Malak Alkhorasani, and Hamzah Luqman. Crosshallu: Do hallucination signals generalize across languages and domains in large language model's internals? *CoRR*, abs/2607.04029, 2026. doi: 10.48550/ARXIV.2607.04029. URL https://doi.org/10.48550/arXiv.2607.04029.

Amos Azaria and Tom M. Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, volume EMNLP 2023 of *Findings of ACL*, pp. 967–976. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.68. URL https://doi.org/10.18653/v1/2023.findings-emnlp.68.

Guy Bar-Shalom, Fabrizio Frasca, Yaniv Galron, Yftah Ziser, and Haggai Maron. Beyond token probes: Hallucination detection via activation tensors with act-vit. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/7b8694d58c34b9bec9c2f29735c3a250-Abstract-Conference.html.

Alexandra Bazarova, Andrei Volodichev, Aleksandr Yugay, Andrey Shulga, Alina Ermilova, Konstantin Polev, Julia Belikova, Rauf Parchiev, Dmitry Simakov, Maxim Savchenko, Andrey V. Savchenko, Serguei Barannikov, and Alexey Zaytsev. Hallucination detection in llms with topological divergence on attention graphs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 15449–15470. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.704. URL https://doi.org/10.18653/v1/2026.acl-long.704.

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Kajdanowicz. Hallucination detection in llms using spectral features of attention maps. In Christos Christodoulopoulos,

Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 24354–24385. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1239. URL https://doi.org/10.18653/v1/2025.emnlp-main.1239.

Anwoy Chatterjee, Yash Goel, and Tanmoy Chakraborty. HIDE and seek: Detecting hallucinations in language models via decoupled representations. *CoRR*, abs/2506.17748, 2025. doi: 10.48550/ARXIV.2506.17748. URL https://doi.org/10.48550/arXiv.2506.17748.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024a. URL https://openreview.net/forum?id=Zj12nzlQbz.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2292–2307, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.146. URL https://aclanthology.org/2022.emnlp-main.146/.

Kedi Chen, Qin Chen, Jie Zhou, He Yishen, and Liang He. DiaHalu: A dialogue-level hallucination evaluation benchmark for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9057–9079, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.529. URL https://aclanthology.org/2024.findings-emnlp.529/.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 1419–1436. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.84. URL https://doi.org/10.18653/v1/2024.emnlp-main.84.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Chengfeng Dou, Ying Zhang, Yanyuan Chen, Zhi Jin, Wenpin Jiao, Haiyan Zhao, Yongqiang Zhao, Zhenwei Tao, and Yun Huang. Detection, diagnosis, and explanation: A benchmark for Chinese medical hallucination evaluation. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4784–4794, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.428/.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. Fact-checking the output of large language models via token-level uncertainty quantification. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9367–9385, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.558. URL https://aclanthology.org/2024.findings-acl.558/.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nat.*, 630(8017):625–630, 2024. doi: 10.1038/S41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

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.

Jiatong Han, Neil Band, Muhammed Razzak, Jannik Kossen, Tim G. J. Rudner, and Yarin Gal. Simple factuality probes detect hallucinations in long-form natural language generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 16209–16226. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-EMNLP.880. URL https://doi.org/10.18653/v1/2025.findings-emnlp.880.

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 Trans. Inf. Syst.*, 43(2):42:1–42:55, 2025. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

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 Comput. Surv.*, 55(12):248:1–248:38, 2023. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

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

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-1147. URL https://doi.org/10.18653/v1/P17-1147.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth A. Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. *CoRR*, abs/2406.15927, 2024. doi: 10.48550/ARXIV.2406.15927. URL https://doi.org/10.48550/arXiv.2406.15927.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466, 2019. doi: 10.1162/TACL_A_00276. URL https://doi.org/10.1162/tacl_a_00276.

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 Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 6449–6464. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.397. URL https://doi.org/10.18653/v1/2023.emnlp-main.397.

Lin Li, Georgia Channing, Suhaas M Bhat, Gabriel Davis Jones, and Yarin Gal. Building reliable long-form generation via hallucination rejection sampling, 2026. URL https://arxiv.org/abs/2606.03628.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.229. URL https://doi.org/10.18653/v1/2022.acl-long.229.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Siyi Liu, Kishaloy Halder, Zheng Qi, Wei Xiao, Nikolaos Pappas, Phu Mon Htut, Neha Anna John, Yassine Benajiba, and Dan Roth. Towards long context hallucination detection. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7842–7850, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.436. URL https://aclanthology.org/2025.findings-naacl.436/.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. A token-level reference-free hallucination detection benchmark for free-form text generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6723–6737, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.464. URL https://aclanthology.org/2022.acl-long.464/.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 9004–9017. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.557. URL https://doi.org/10.18653/v1/2023.emnlp-main.557.

Andreas Marfurt and James Henderson. Unsupervised token-level hallucination detection from summary generation by-products. In Antoine Bosselut, Khyathi Chandu, Kaustubh Dhole, Varun Gangal, Sebastian Gehrmann, Yacine Jernite, Jekaterina Novikova, and Laura Perez-Beltrachini (eds.), *Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pp. 248–261, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.gem-1.21. URL https://aclanthology.org/2022.gem-1.21/.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173/.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pp. 10862–10878, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.585. URL https://aclanthology.org/2024.acl-long.585/.

Mengjia Niu, Hamed Haddadi, and Guansong Pang. Robust hallucination detection in llms via adaptive token selection. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/b7c43d4a79dede363a2d061c6158e5a5-Abstract-Conference.html.

Yuya Ogasa and Yuki Arase. Hallucinated span detection with multi-view attention features. In Lea Frermann and Mark Stevenson (eds.), *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pp. 381–394, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-340-1. doi: 10.18653/v1/2025.starsem-1.31. URL https://aclanthology.org/2025.starsem-1.31/.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

Hongxing Pan, Yingying Guo, Wenqing Kuang, and Jiashi Lu. Mind the unseen mass: Unmasking llm hallucinations via soft-hybrid alphabet estimation, 2026. URL https://arxiv.org/abs/2604.19162.

Siya Qi, Yudong Chen, Runcong Zhao, Qinglin Zhu, Zhanghao Hu, Wei Liu, Yulan He, Zheng Yuan, and Lin Gui. Detecting contextual hallucinations in large language models with frequency-aware attention. In *Forty-third International Conference on Machine Learning*.

Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomas Cerny. Detecting hallucinations in large language model generation: A token probability approach, 2024. URL https://arxiv.org/abs/2405.19648.

Shreyas N. Samaga, Gilberto Gonzalez Arroyo, and Tamal K. Dey. Halluzig: Hallucination detection using zigzag persistence. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2026 - Volume 1: Long Papers, Rabat, Morocco, March 24-29, 2026*, pp. 3466–3482. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.EACL-LONG.159. URL https://doi.org/10.18653/v1/2026.eacl-long.159.

Ponhvoan Srey, Xiaobao Wu, Cong-Duy T. Nguyen, and Anh Tuan Luu. Learning uncertainty from sequential internal dispersion in large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 40088–40106. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.1862. URL https://doi.org/10.18653/v1/2026.acl-long.1862.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14379–14391, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.854. URL https://aclanthology.org/2024.findings-acl.854/.

Yufei Tao, Adam Hiatt, Erik Haake, Antonie J. Jetter, and Ameeta Agrawal. When context leads but parametric memory follows in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4034–4058, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.234. URL https://aclanthology.org/2024.emnlp-main.234/.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025. doi: 10.48550/ARXIV.2507.06261. URL https://doi.org/10.48550/arXiv.2507.06261.

Gijs van Dijk. Detecting hallucinations in large language models via internal attention divergence signals. In T. Y. S. S. Santosh, Juan Diego Rodriguez, and Ona de Gibert (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 60–75. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-SRW.6. URL https://doi.org/10.18653/v1/2026.acl-srw.6.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. Factuality of large language models: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 19519–19529. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.1088. URL https://doi.org/10.18653/v1/2024.emnlp-main.1088.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1259. URL https://doi.org/10.18653/v1/d18-1259.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How language model hallucinations can snowball. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59670–59684. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/zhang24ay.html.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219, 2023. doi: 10.48550/ARXIV.2309.01219. URL https://doi.org/10.48550/arXiv.2309.01219.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. ICR probe: Tracking hidden state dynamics for reliable hallucination detection in llms. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 17986–18002. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.880. URL https://doi.org/10.18653/v1/2025.acl-long.880.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14544–14556, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.968. URL https://aclanthology.org/2023.findings-emnlp.968/.

Supplementary Material

A Experimental Setup 18

A.1 Evaluation Protocol and Implementation 18

A.2 Baselines 18

B Method Details 19

B.1 End-to-End Algorithm 19

C Additional Quantitative Results 19

C.1 Length-Matched Evaluation 19

C.2 Multiple-Generation Evaluation 20

C.3 Ablation Studies 21

C.4 Cross-Dataset Transfer Experiment 21

C.5 Sensitivity Analysis 21

D Diagnostic Analyses 22

D.1 Order-Destruction Analysis 22

D.2 Evolution of Detection Signals during Generation 24

D.3 Case Studies 25

E Proofs for the Theoretical Analysis of DOOR 26

E.1 Proof of Proposition 3.1 26

E.2 Proof of Proposition 3.2 27

E.3 Proof of Theorem 3.3 28

E.4 Scope of the Guarantees 29

A Experimental Setup

A.1 Evaluation Protocol and Implementation

Models and Datasets. We evaluate Qwen3-8B(Yang et al., 2025), Llama-3.1-8B-Instruct(Grattafiori et al., 2024), and Mistral-7B-Instruct(Jiang et al., 2023) on TruthfulQA(Lin et al., 2022), TriviaQA(Joshi et al., 2017), NQ-Open(Kwiatkowski et al., 2019), HotpotQA(Yang et al., 2018), GSM8K(Cobbe et al., 2021), and HaluEval(Li et al., 2023). For each split seed and backbone, the evaluation pool contains 500 TruthfulQA responses, 3,000 responses from each of TriviaQA, NQ-Open, and HotpotQA, and 1,000 responses from each of GSM8K and HaluEval. Cross-dataset transfer uses the four factual QA datasets and excludes GSM8K and HaluEval. Each model produces one greedy response per question with at most 512 new tokens and a maximum prompt length of 1024.

Data Splits and Evaluation Metrics. We construct three question-disjoint train/validation/test splits with seeds 1701/1702/1703, giving an approximate 56/14/30 partition. Models are trained on the training split, validation is used only when required by a baseline, and the test split is evaluated once. We report AUROC and AUPRC as mean $\pm$ standard deviation over the three splits. Incorrect responses form the positive class. Task-specific exact matching uses the extracted final number for GSM8K and normalized textual matching for the other benchmarks.

Label Generation. Our label-generation code is adapted from the publicly available implementation of Orgad et al. (2025). Following established practice in representation-level hallucination detection, we derive response-level correctness labels automatically from generated responses and reference answers (Orgad et al., 2025; Bar-Shalom et al., 2025; Zhang et al., 2025; Niu et al., 2025; Chuang et al., 2024). For GSM8K, correctness is determined exclusively by exact matching of the extracted final numerical answer. For short-answer QA benchmarks, including TriviaQA, NQ-Open, and HotpotQA, correctness is determined by normalized matching against the gold answer or one of its annotated aliases within the generated response. For benchmarks with longer or more variable natural-language answers, a response is labelled faithful if normalized exact matching succeeds, ROUGE-L F1 exceeds 0.7, or BLEURT-20 exceeds 0.5. For TruthfulQA, the response must additionally not match an incorrect reference more strongly than a correct one. For HaluEval, we apply the same rule using the official correct answer as the reference; the dataset-provided hallucinated answer is not used as an incorrect reference. All compared detectors use exactly the same generated responses and correctness labels.

Implementation Details. Experiments are implemented in PyTorch and Transformers and run on NVIDIA H100 80 GB GPUs. Generation uses SDPA, while attention-based feature extraction uses a single eager forward pass over the stored response.

A.2 Baselines

We compare `DOOR` with representative attention-based, spectral, and hidden-state hallucination detectors. These baselines cover complementary signals from the generation process, including attention allocation, frequency-domain structure, and token-level internal representations. For clarity, we summarize the core signal used by each method.

- **Lookback Lens.** Measures the extent to which each generated token attends back to the input prompt rather than to previously generated tokens (Chuang et al., 2024). This signal characterizes whether generation remains anchored to the provided context throughout decoding.
- **AttentionDvg.** Uses the concentration or divergence of the attention distribution as a confidence-related signal (van Dijk, 2026). More focused attention patterns are interpreted as reflecting more decisive information selection, whereas diffuse patterns may indicate uncertain or less grounded generation.
- **FourierAttention.** Transforms attention patterns into the frequency domain and characterizes them through high-frequency spectral energy (Qi et al.). The method captures local fluctuations and structural irregularities in attention that may be obscured in the original attention space.

- **LapEigVals.** Constructs a graph from causal attention relations and represents its global structure using leading graph-Laplacian eigenvalues (Binkowski et al., 2025). This spectral representation summarizes how information-flow patterns are organized across the generated sequence.
- **ICR Probe.** Aggregates internal confidence-reflection scores across response tokens and transformer layers (Zhang et al., 2025). It exploits the hypothesis that hidden representations encode signals about whether the model is confident in, or uncertain about, its ongoing generation.
- **LLMs Know.** Detects hallucination from the hidden state of the answer token (Orgad et al., 2025), exploiting the concentration of truthfulness-related information in answer tokens.
- **ACT-ViT.** Treats the response-wise hidden-state trajectory as a structured sequence and applies a vision-transformer readout to its pooled representation (Bar-Shalom et al., 2025). It therefore captures interactions among hidden states across multiple generated positions rather than relying on a single token.
- **HaMI.** Uses multiple-instance learning to aggregate hallucination evidence from token-level hidden features (Niu et al., 2025). The approach is designed for responses in which only a small subset of generated tokens may carry strong hallucination-related signals.

B Method Details

B.1 End-to-End Algorithm

Algorithm 1 gives the complete training and inference procedure. The same feature extractor and fixed layer–head ordering are used for every example; only the feature standardizer and linear probe are learned.

Algorithm 1 End-to-end training and inference with `DOOR`

Require: Frozen language model f; labeled training set $\mathcal{D}_{\text{tr}} = \{(x_i, o_i, z_i)\}_{i=1}^{N}$; profile length K
Ensure: Training statistics $(\widehat{\mathbf{m}}, \widehat{\mathbf{s}})$ and probe parameters $(\mathbf{w}, b)$

function DoorFeature(x, o)
  Apply the chat template and run one teacher-forced pass on $[x; o]$
  Extract generated-to-prompt attention blocks $\mathbf{P}^{\ell,h}$ (Section 3.2) // Attention Extraction
  for all layer–head pairs (ℓ, h) **do**
    Compute reuse intensity $\boldsymbol{\mu}^{\ell,h}$ and dispersion $\mathbf{c}^{\ell,h}$ using Eqs. (1)–(2) // Prompt-Reuse Profiling
    Apply Π_K at the ranks in Eq. (3) and form the RI–RD head profile (Section 3.3)
  end for
  return concatenated layer–head representation $\mathbf{x}_{\text{RD+RI}}$
end function
Training:
for $i = 1, \ldots, N$ **do**
  $\mathbf{x}_i \leftarrow$ DoorFeature(x_i, o_i)
end for
Standardize features using training-set statistics only
Fit the class-balanced ℓ_2-regularized logistic probe (Section 3.4) // Linear Probe Training
Inference:
function Score(x, o)
  $\mathbf{x} \leftarrow$ DoorFeature(x, o)
  Standardize $\mathbf{x}$ using stored training statistics
  return return the probe output as the hallucination score // Hallucination Scoring
end function

C Additional Quantitative Results

C.1 Length-Matched Evaluation

Although rank-based profiling maps variable-length prompt descriptors into a fixed-dimensional representation, the underlying attention statistics may still vary with prompt and response length.

Table 3: Length-matched evaluation of DOOR on Qwen3-8B. Hallucinated and faithful responses are matched by prompt and response length, while the detector remains fixed. We report AUROC (%, mean ± standard deviation) over three split seeds.

Setting	TruthfulQA	TriviaQA	NQ-Open	HotpotQA	GSM8K	Avg.
Original	80.8 ±2.0	85.6 ±0.6	75.2 ±1.5	85.8 ±1.5	79.1 ±5.0	81.3 ±1.2
Length-matched	82.1 ±1.7	86.6 ±2.2	75.5 ±0.9	84.7 ±4.0	76.9 ±3.1	81.2 ±2.2

Table 4: Within-question AUROC (%) of DOOR on Qwen3-8B across five benchmarks. For each question, we sample five responses and compute AUROC only when both correct and incorrect responses are present. Question-level AUROCs are averaged equally within each split. Each benchmark reports mean ± standard deviation over three question-disjoint split seeds (2701, 2702, and 2703). For Avg., we first average the five benchmark AUROCs within each seed, then report the mean ± standard deviation of the three seed-level averages.

Method	TruthfulQA	TriviaQA	NQ-Open	HotpotQA	GSM8K	Avg.
DOOR	66.9 ± 2.6	74.2 ± 1.7	72.3 ± 0.6	75.4 ± 3.3	81.2 ± 0.9	74.0 ± 1.8

We therefore conduct a length-matched evaluation to test whether the performance of DOOR can be explained primarily by these length-related factors.

For each dataset and split seed, we construct a matched subset from the original test set. Each example is represented by its prompt length P and response length T, using $(\log(1+P), \log(1+T))$ after standardizing the two dimensions. Incorrect and faithful responses are then paired one-to-one with examples of similar length using nearest-neighbor matching. The DOOR representation, feature standardizer, and linear probe remain exactly those learned from the original training split; no retraining or adaptation is performed after matching.

Table 3 compares the original and length-matched evaluation on Qwen3-8B. The average AUROC remains nearly unchanged, from 81.3 to 81.2, after controlling for prompt and response length. Performance is preserved or slightly improved on TruthfulQA, TriviaQA, and NQ-Open, remains close to the original result on HotpotQA, and decreases modestly on GSM8K. These results indicate that the predictive performance of DOOR cannot be explained primarily by prompt or response length.

C.2 Multiple-Generation Evaluation

The standard evaluation compares responses to different questions and may therefore reflect differences in prompt-level difficulty. To isolate response-specific discrimination, we evaluate whether DOOR assigns a higher hallucination score to an incorrect response than to a correct response generated for the same question.

We repeat the question-disjoint split protocol with seeds 2701, 2702, and 2703. For each test question, we sample $N = 5$ responses from Qwen3-8B with temperature 1.0 and top-p 0.95, keeping the prompt fixed. Each response is labeled using the correctness criteria in Appendix A.1. For each split, the detector is trained using the original single-response training procedure and kept frozen during this evaluation. The additional sampled responses are used only for testing. For each question containing both correct and incorrect sampled responses, we compute a within-question AUROC from all incorrect–correct response pairs, counting tied scores as one-half. We average the question-level AUROCs equally within each split. For each benchmark, we report the mean and standard deviation of this value over the three split seeds. To obtain Avg., we first average the five benchmark AUROCs within each seed and then compute the mean and standard deviation across seeds. Questions whose sampled responses all have the same label do not contribute to the within-question AUROC.

Table 4 reports within-question AUROC on five benchmarks. DOOR achieves 66.9% on TruthfulQA, 74.2% on TriviaQA, 72.3% on NQ-Open, 75.4% on HotpotQA, and 81.2% on GSM8K, with an average of 74.0%. Because every AUROC comparison is between responses to the same question, these results provide evidence of response-specific predictive information beyond differences in difficulty between questions.

Table 5: Comparison of controlled variants of DOOR. We report AUROC and AUPRC (%) as mean ± standard deviation over three question-disjoint split seeds. Dark and light blue cells indicate the best and second-best results, respectively, within each backbone–dataset setting.

Backbone	Variant	TruthfulQA		TriviaQA		NQ-Open		HotpotQA		GSM8K		Avg.	
		AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Qwen3-8B	V1	79.8 ±2.8	94.9 ±1.0	84.6 ±1.0	77.9 ±1.5	73.9 ±2.6	79.0 ±3.4	84.8 ±0.7	92.1 ±2.6	77.9 ±3.3	89.1 ±3.6	80.2 ±1.0	86.6 ±1.2
	V2	73.2 ±2.2	91.7 ±2.6	83.6 ±0.9	76.6 ±2.6	73.0 ±1.2	79.9 ±1.7	83.2 ±2.2	90.4 ±3.3	76.7 ±4.5	88.7 ±1.7	77.9 ±1.1	85.5 ±1.1
	V3	76.8 ±2.1	93.6 ±1.2	83.3 ±1.8	75.8 ±2.8	**75.2** ±0.6	77.5 ±2.0	83.4 ±1.9	90.7 ±3.4	76.9 ±4.9	88.8 ±2.0	79.1 ±1.2	85.3 ±1.1
	V4	75.1 ±3.1	92.8 ±1.2	79.4 ±2.7	76.8 ±1.5	69.3 ±2.2	77.5 ±1.8	79.0 ±3.4	90.1 ±2.0	73.2 ±2.9	86.9 ±1.6	75.2 ±2.5	84.8 ±1.4
	V5	78.4 ±2.2	94.1 ±1.3	84.8 ±2.1	78.3 ±3.7	74.8 ±2.1	**80.3** ±2.6	85.1 ±2.0	92.0 ±3.8	76.8 ±4.1	87.4 ±2.8	80.0 ±1.2	86.4 ±1.3
	V6	72.0 ±3.0	92.8 ±0.8	84.8 ±3.3	77.3 ±7.4	73.2 ±0.8	79.3 ±0.9	80.1 ±0.8	87.4 ±1.3	76.2 ±2.0	89.2 ±1.5	77.3 ±1.0	85.2 ±1.6
	DOOR	**80.8** ±2.0	**95.0** ±0.9	**85.6** ±0.6	**79.0** ±1.3	**75.2** ±1.5	80.1 ±2.5	**85.8** ±1.5	**92.3** ±3.1	**79.1** ±5.0	**89.7** ±3.5	**81.3** ±1.2	**87.2** ±1.1
Llama-3.1-8B	V1	75.2 ±1.6	91.7 ±0.2	89.3 ±2.9	83.3 ±4.1	67.3 ±1.1	64.8 ±1.7	90.4 ±2.9	97.7 ±0.6	81.1 ±2.6	58.6 ±6.0	80.7 ±1.0	79.2 ±1.5
	V2	74.5 ±7.1	91.2 ±2.6	89.0 ±1.6	84.5 ±5.6	69.5 ±2.0	66.2 ±3.3	89.3 ±1.8	97.4 ±0.1	79.9 ±4.8	57.3 ±12.2	80.4 ±1.8	79.3 ±2.8
	V3	75.8 ±5.3	91.5 ±2.8	90.3 ±2.2	84.6 ±4.9	69.2 ±1.7	67.0 ±1.7	89.9 ±1.8	97.7 ±0.4	80.1 ±4.7	58.3 ±11.8	81.1 ±1.6	79.8 ±2.6
	V4	70.3 ±2.8	90.2 ±1.3	83.5 ±3.0	82.7 ±2.2	63.0 ±1.9	66.2 ±1.1	84.4 ±2.7	95.4 ±1.0	77.0 ±2.5	63.1 ±2.8	75.6 ±2.1	79.5 ±1.7
	V5	72.4 ±6.7	89.8 ±1.8	89.9 ±1.5	84.4 ±3.3	69.4 ±1.8	68.2 ±1.4	**90.9** ±2.0	97.2 ±0.3	80.3 ±3.5	58.6 ±8.8	80.6 ±1.6	79.6 ±1.9
	V6	72.7 ±1.8	91.7 ±0.6	87.1 ±1.7	84.3 ±4.7	67.9 ±2.0	66.7 ±1.1	89.1 ±1.9	97.2 ±0.1	**84.3** ±0.7	**66.3** ±3.3	80.2 ±0.8	81.2 ±1.2
	DOOR	**76.7** ±4.7	**92.4** ±1.5	**90.4** ±2.3	**85.2** ±4.1	**69.9** ±0.7	**68.3** ±0.4	90.5 ±2.2	**97.8** ±0.4	84.2 ±3.3	**66.3** ±5.3	**82.3** ±1.3	**82.0** ±1.4
Mistral-7B	V1	78.7 ±3.7	93.7 ±1.2	77.9 ±2.0	65.1 ±4.0	68.6 ±1.4	69.5 ±1.3	71.8 ±0.4	81.2 ±2.5	80.8 ±1.7	85.0 ±3.2	75.6 ±1.0	78.9 ±1.2
	V2	74.0 ±7.7	93.1 ±1.6	77.7 ±1.4	65.7 ±1.0	68.1 ±1.4	71.2 ±1.4	**75.0** ±2.4	82.8 ±1.5	81.3 ±3.2	85.5 ±0.8	75.2 ±1.8	79.7 ±0.6
	V3	77.3 ±7.2	93.7 ±1.7	78.2 ±0.9	65.8 ±1.6	**69.5** ±1.4	71.7 ±2.1	74.4 ±4.2	82.8 ±0.8	81.6 ±2.7	85.7 ±0.2	76.2 ±1.8	79.9 ±0.6
	V4	72.0 ±3.6	91.8 ±1.0	71.5 ±2.4	63.8 ±1.3	62.2 ±2.7	69.8 ±1.2	68.8 ±3.1	80.5 ±1.4	74.6 ±2.6	83.8 ±1.8	69.8 ±2.3	77.9 ±1.1
	V5	77.5 ±2.9	93.4 ±0.8	78.3 ±0.7	**66.7** ±1.6	68.4 ±0.4	70.7 ±1.7	74.5 ±2.5	82.7 ±1.3	**81.7** ±0.5	85.6 ±0.9	76.1 ±0.8	79.8 ±0.6
	V6	68.1 ±4.7	90.7 ±1.0	76.1 ±1.1	62.0 ±1.7	69.3 ±2.6	71.4 ±2.4	74.5 ±3.2	82.9 ±1.8	80.3 ±1.1	84.8 ±3.4	73.7 ±1.3	78.4 ±1.0
	DOOR	**78.8** ±4.9	**94.2** ±0.4	**78.6** ±0.6	66.0 ±1.0	**69.5** ±1.3	**72.0** ±1.8	**75.0** ±1.8	**83.0** ±0.5	**81.7** ±1.2	**86.1** ±1.5	**76.7** ±1.1	**80.3** ±0.5

C.3 Ablation Studies

All controlled variants use the same data splits, labels, feature standardization, and linear probe; profile-based variants also use the same profile length and stabilizer. V1 retains only the reuse-dispersion (RD) channel, while V2 retains only the reuse-intensity (RI) channel. V3 replaces the relative-dispersion coefficient of variation with the raw standard deviation. V4 replaces each RI and RD rank profile with its mean and standard deviation. V5 preserves cross-channel token correspondence by sorting prompt positions according to RI and sampling RD at the same RI-induced ranks. This variant has the same feature dimension as the independently ranked DOOR representation. V6 computes both profiles from generated-to-generated attention instead of generated-to-prompt attention. For the causal generated-to-generated attention block, all masked entries are excluded from the computation of the reuse statistics rather than treated as zero-valued attention entries. Thus, the RI and RD descriptors are computed only over valid causal attention entries. These variants isolate channel complementarity, dispersion normalization, distributional profiling, rank coupling, and the choice of attention block.

C.4 Cross-Dataset Transfer Experiment

Figure 3 is restricted to Qwen3-8B and the four factual QA suites. GSM8K and HaluEval are used only in in-distribution tables (Tables 1 and 5). Methods are Lookback Lens, AttentionDvg, FourierAttention, and locked DOOR. For each seed, the detector is fit on the source-dataset *train* fold. If source = target (diagonal), we score that fold's *test* split (in-distribution). If source ≠ target (off-diagonal), we score *all* target examples in the corresponding per-seed pool—500 for TruthfulQA and 3,000 each for TriviaQA, NQ-Open, and HotpotQA (out-of-distribution). We report AUROC mean±std over seeds 1701/1702/1703. Row-mean OOD averages the three off-diagonal targets. ID in the table header is the mean of the four diagonal cells; OOD is the mean of the twelve off-diagonal cells.

C.5 Sensitivity Analysis

We examine the sensitivity of DOOR to two design choices: the rank-profile length K and the dispersion stabilizer ε. All experiments reuse the Qwen3-8B evaluation pools, grouped splits, labels, and linear-probe protocol defined in Appendix A.1. Unless explicitly varied, we keep the main-experiment settings fixed at $K = 8$ and $\varepsilon = 10^{-8}$. The analysis uses the five original evaluation suites and excludes HaluEval.

Profile length. Table 6 varies K while holding $\varepsilon = 10^{-8}$ fixed. Performance improves rapidly with a small number of sampled ranks and then largely saturates, indicating that a compact rank profile is

Table 6: Sensitivity of DOOR to the profile length K. We report AUROC (%, mean $\pm$ standard deviation) over three question-disjoint split seeds. The blue row denotes the locked setting $K = 8$.

K	TruthfulQA	TriviaQA	NQ-Open	HotpotQA	GSM8K
2	75.0 $\pm$1.1	83.5 $\pm$0.3	75.8 $\pm$0.7	83.0 $\pm$1.0	80.3 $\pm$5.1
4	77.5 $\pm$2.6	85.3 $\pm$0.9	76.2 $\pm$1.0	84.2 $\pm$1.7	80.8 $\pm$5.0
8	80.8 $\pm$2.0	85.6 $\pm$0.6	75.2 $\pm$1.5	85.8 $\pm$1.5	79.1 $\pm$5.0
16	81.4 $\pm$2.2	85.2 $\pm$0.4	75.1 $\pm$1.7	86.0 $\pm$1.7	77.8 $\pm$4.9
32	81.5 $\pm$2.4	85.1 $\pm$0.6	75.2 $\pm$1.7	86.8 $\pm$1.8	76.0 $\pm$5.0

Table 7: Sensitivity of DOOR to the dispersion stabilizer ε. We report AUROC (%, mean $\pm$ standard deviation) over three question-disjoint split seeds. The blue row denotes the locked setting $\varepsilon = 10^{-8}$.

ε	TruthfulQA	TriviaQA	NQ-Open	HotpotQA	GSM8K
10^{-12}	80.7 $\pm$2.4	85.6 $\pm$0.6	75.4 $\pm$1.4	85.7 $\pm$1.5	79.3 $\pm$4.9
$\mathbf{10^{-8}}$	80.8 $\pm$2.0	85.6 $\pm$0.6	75.2 $\pm$1.5	85.8 $\pm$1.5	79.1 $\pm$5.0
10^{-6}	81.1 $\pm$1.7	85.6 $\pm$0.4	75.6 $\pm$1.7	85.5 $\pm$1.5	79.0 $\pm$5.2
10^{-4}	81.2 $\pm$1.7	85.0 $\pm$1.1	76.4 $\pm$1.2	85.3 $\pm$1.3	76.4 $\pm$4.2
10^{-3}	79.7 $\pm$2.5	83.4 $\pm$1.2	76.8 $\pm$0.2	85.6 $\pm$1.8	75.2 $\pm$3.9
10^{-2}	76.5 $\pm$3.1	83.0 $\pm$1.6	76.2 $\pm$1.0	83.9 $\pm$2.4	75.3 $\pm$3.9

sufficient to preserve the useful distributional structure. The locked choice $K = 8$ lies within this plateau rather than at an isolated optimum.

Dispersion stabilizer. Table 7 varies ε while holding $K = 8$ fixed. Performance is stable across a broad range of small stabilizers and degrades only when ε becomes sufficiently large to suppress relative-dispersion variation. This supports the locked setting $\varepsilon = 10^{-8}$ as a numerically stable choice rather than a finely tuned hyperparameter.

D Diagnostic Analyses

D.1 Temporal order Analysis

Locked DOOR is permutation-invariant with respect to generation steps. We therefore use this analysis to test whether representative order-sensitive detectors rely on the ordering assumed by their features. For SIVR (Srey et al., 2026) and CrossHallu (Alansari et al., 2026), we permute the generation-step axis. For FourierAttention (Qi et al.), we permute the attended-position axis along which its spectrum is computed. These interventions preserve the underlying multiset of values while destroying the ordering used by each method.

Specifically, SIVR processes an ordered sequence of token-dispersion statistics; CrossHallu uses the consecutive confidence gradient $G_{\text{true}} = \text{mean}_t|p_{t+1} - p_t|$; and FourierAttention measures high-frequency attention energy E_H. Each detector is trained once on the original, unpermuted training data. At evaluation time, all learned parameters and preprocessing statistics are frozen. For each test example, we randomly permute the relevant axis, recompute the corresponding order-sensitive feature, and evaluate the unchanged detector. No retraining, fine-tuning, or recalibration is performed after permutation. We interpret this intervention as a test of reliance on the explicit ordering used by these detectors, rather than as evidence that autoregressive generation order is universally irrelevant.

Experiment Protocols. All methods use the same 2×5 grid: Qwen3-8B and Mistral-7B across the five original in-distribution datasets (TruthfulQA, TriviaQA, NQ-Open, HotpotQA, and GSM8K), with the dataset-specific per-seed sample sizes, automatic labels, and grouped splits defined in Appendix A.1. The diagnostic experiments use independently sampled question sets from those used in the main comparison in Table 1. Therefore, the "Original" values reported in Table 8 are the unpermuted references for the diagnostic subsets and need not numerically match the corresponding results in Table 1. Within each model–dataset setting, the Original and Shuffle conditions are evaluated on exactly the same examples. For each example we draw R=10 deterministic permutations.

Main Results of Order Destruction Experiments. The per-cell results in Table 8 yield mean changes of $+0.76$, $+0.19$, and -0.15 AUROC points for SIVR, CrossHallu, and FourierAttention, respectively. These averages do not show a consistent performance drop after the tested permutations, although individual effects vary by method and dataset. For example, CrossHallu drops by 4.8 points

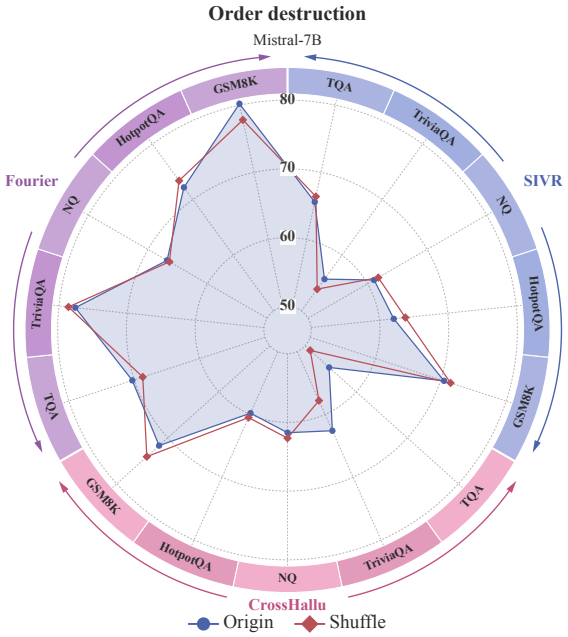


Figure 6: Order-destruction results on Mistral-7B. Radial coordinates report AUROC (%) for SIVR, CrossHallu, and FourierAttention across the five evaluation datasets. Blue circles denote the original features and red diamonds denote features after destroying the method-specific ordering while preserving the underlying values. The close correspondence between the two traces illustrates the limited effect of order destruction; complete per-cell results are reported in Table 8.

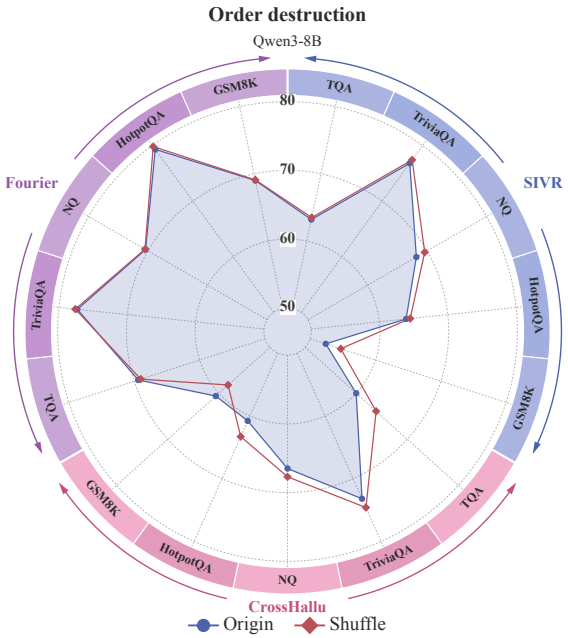


Figure 7: Effect of order destruction across five datasets on Qwen3-8B. AUROC (%) under original vs. shuffled token order.

Table 8: Effect of destroying the native ordering used by each detector across all 2×5 model–dataset cells. Original uses the native order-sensitive feature and Shuffle destroys the corresponding ordering axis while preserving the underlying values. AUROC is reported in percent.

Detector	Backbone	TruthfulQA		TriviaQA		NQ-Open		HotpotQA		GSM8K	
		Orig.	Shuf.	Orig.	Shuf.	Orig.	Shuf.	Orig.	Shuf.	Orig.	Shuf.
SIVR	Qwen3-8B	63.2	63.5	76.8	77.4	68.2	69.6	63.9	64.5	52.4	54.7
	Mistral-7B	65.6	66.4	55.7	53.9	61.1	61.8	62.1	63.8	70.5	71.5
CrossHallu	Qwen3-8B	60.0	63.9	73.2	74.6	66.5	67.7	60.8	63.3	60.6	58.2
	Mistral-7B	54.7	51.0	62.6	57.8	61.5	62.3	59.8	60.5	71.7	74.1
Fourier	Qwen3-8B	69.4	69.0	77.3	77.6	70.4	70.5	79.3	79.8	69.0	69.1
	Mistral-7B	70.3	68.7	77.6	78.6	66.8	66.4	72.2	73.4	80.2	77.8

on TriviaQA with Mistral-7B. Thus, the experiment suggests limited reliance on the tested ordering in aggregate, without implying that token order is generally irrelevant.

D.2 Evolution of Detection Signals during Generation

Finding 2 examines how much hallucination-discriminative information is already linearly accessible from the prompt and how this signal evolves as generated tokens are progressively incorporated into the representation. The goal is to characterize where detectable signals are available throughout generation rather than to introduce a new detector.

Representation and Protocol. Let $h_j^{(\ell)}$ denote the residual-stream hidden state at transformer block ℓ and token position j. Let P denote the prompt tokens and $G_{\le\tau}$ the first $\lfloor \tau T \rfloor$ generated tokens, where $\tau \in \{0, 0.10, 0.25, 0.50, 1.00\}$ and T is the full response length. At generation progress τ, we define the available prefix as $C_\tau = [P; G_{\le\tau}]$ and construct the mean-pooled representation

$$z_\tau^{(\ell)} = \mathrm{mean}_{j \in C_\tau} h_j^{(\ell)}.$$

Thus, $\tau = 0$ contains prompt tokens only, while later values progressively incorporate generated-token hidden states into the same pooling window. All experiments use the fixed middle transformer block $\ell = L/2$. A separate probe is trained independently at each generation-progress value, with no parameter sharing across progress. Results are reported separately for each dataset rather than averaged across datasets.

Probe Configurations. The main linear probe consists of `StandardScaler` followed by a class-balanced `LogisticRegression` classifier (`liblinear`, `max_iter=2000`), trained on the training split and evaluated on the test split. As a nonlinear robustness check, we also use an MLP with architecture $d\to256\to1$, GELU activation, and dropout 0.1. The MLP is trained for 30 epochs with positive-class-weighted binary cross-entropy, and the checkpoint with the highest validation AUROC is used for test evaluation. Hidden states are extracted in float16 and converted to float32 before probe training.

Generation-Progress Results. Figure 8 reports results separately for each dataset. A high AUROC at 0% shows that substantial hallucination-discriminative information is already accessible from prompt-token hidden states. As generation proceeds, the trajectories are strongly task-dependent rather than uniformly increasing. In particular, generated-token representations provide clear additional information in some reasoning settings, especially GSM8K, whereas several factual and multi-hop QA settings exhibit flat or non-monotonic trajectories. These results indicate that additional generated tokens do not yield uniformly cumulative gains in hallucination detection.

Because each generated-token hidden state is itself conditioned on the prompt through causal attention, improvements after 0% should not be interpreted as prompt-independent evidence. They instead indicate that additional hallucination-discriminative information becomes accessible through generation-side representations in some settings.

Separated-Pooling Control. The joint pooling above changes the relative contribution of prompt states as generated tokens are progressively added. We therefore perform an additional control in which prompt and generated-token hidden states are pooled separately. For generation progress

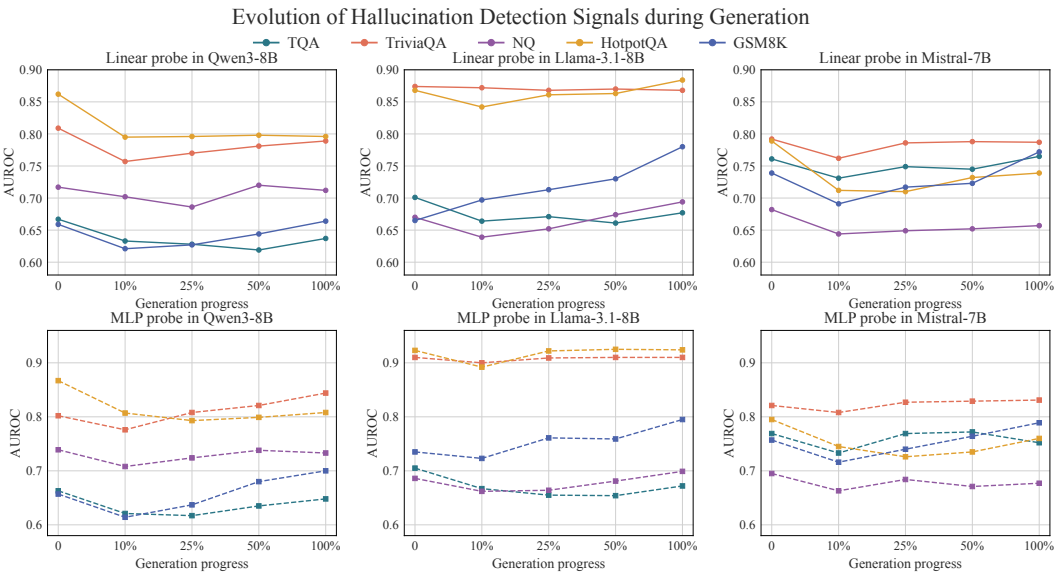


Figure 8: Evolution of hallucination-detection signals during generation. Each panel reports AUROC from prompt-only pooling (0%) to pooling over the full generated response (100%) at the fixed middle layer. The top row uses linear probes and the bottom row uses MLP probes; columns correspond to the three evaluated language models. TQA and NQ denote TruthfulQA and NQ-Open.

Table 9: Generation-progress analysis with separately pooled prompt and generated-token representations. We report AUROC (%) on three representative benchmarks.

Backbone	Dataset	0%	10%	25%	50%	100%
Qwen3-8B	TruthfulQA	66.7	66.8	64.6	64.2	64.0
	HotpotQA	86.2	85.7	80.6	81.6	82.6
	GSM8K	65.9	66.1	66.1	68.9	72.5
Llama-3.1-8B	TruthfulQA	70.1	71.1	71.1	65.8	65.8
	HotpotQA	86.8	86.8	86.9	87.6	86.5
	GSM8K	66.5	70.2	74.8	75.4	81.7

$\tau > 0$, we construct

$$z_\tau^{(\ell)} = \left[\text{mean}_{j \in P}\, h_j^{(\ell)}; \text{mean}_{j \in G_{\leq \tau}}\, h_j^{(\ell)}\right],$$

This construction preserves the prompt summary as a fixed component while allowing generation-side information to be added independently. We use the same fixed middle layer and linear-probe configuration as in the original generation-progress analysis.

Table 9 reports the control on three representative benchmarks spanning factual knowledge (TruthfulQA), multi-hop reasoning (HotpotQA), and mathematical reasoning (GSM8K).The separated-pooling control preserves the same qualitative pattern. TruthfulQA shows no sustained improvement as generation progresses, while HotpotQA remains largely stable or below its prompt-only performance. In contrast, GSM8K benefits from later generation-side representations, with the largest improvement observed for Llama-3.1-8B. Because the prompt summary is preserved as a separate component throughout this control, the non-monotonic behavior observed in the generation-progress analysis cannot be explained solely by dilution of prompt states under joint mean pooling. Overall, the results show that generation-side representations provide task-dependent rather than uniformly cumulative hallucination-discriminative information.

D.3 Case Studies

Attribution Method. Detection remains example-level. For visualization, we decompose the frozen probe logit into contributions from prompt tokens and separately measure the influence of individual generation steps. Consider a channel-specific probe operating on the layer–head descriptor $v^{\ell,h}$, where

$v^{\ell,h} = c^{\ell,h}$ for RD and $v^{\ell,h} = \mu^{\ell,h}$ for RI. For each sampled rank k, the rank-profile interpolation can be written as

$$[\Pi_K(v^{\ell,h})]_k = \sum_{j \in \mathcal{P}} \alpha_{\ell,h,k,j}\, v_j^{\ell,h},$$

where $\alpha_{\ell,h,k,j}$ is the interpolation weight assigned to prompt position j. Rearranging the standardized linear-probe logit then gives

$$s = C + \sum_{j \in \mathcal{P}} A_j, \qquad A_j = \sum_{\ell,h,k} \frac{w_{\ell,h,k}}{s_{\ell,h,k}}\, \alpha_{\ell,h,k,j}\, v_j^{\ell,h},$$

where $w_{\ell,h,k}$ is the probe weight, $s_{\ell,h,k}$ is the training-set standardization scale, and

$$C = b - \sum_{\ell,h,k} \frac{w_{\ell,h,k} m_{\ell,h,k}}{s_{\ell,h,k}}$$

collects the probe intercept and the contribution of the training-set standardization means $m_{\ell,h,k}$. This decomposition reconstructs the original probe logit up to numerical error on the order of 10^{-7}.

To assess the influence of generation step t, let $\mathcal{B} = \{\mathbf{P}^{\ell,h}\}_{\ell,h}$ denote the collection of GP attention blocks across all layers and heads, and let $\mathcal{B}^{-t}$ denote the same collection after removing row t from every block. We define

$$\Delta_t = s(\mathcal{B}) - s(\mathcal{B}^{-t}),$$

where the corresponding RI or RD descriptors are recomputed after row removal. Unlike the prompt-token contributions A_j, the generation-step influences Δ_t do not form an exact additive decomposition of the logit, because removing a generation row changes the column-wise statistics used to construct the descriptors.

Visualization and Case Selection. The main-text figure combines the mean generated-to-prompt attention map with prompt-token attribution A_j and generation-step influence Δ_t. The generated text indicates the semantic hallucination itself, while the attribution colours represent detector evidence rather than token correctness. Panel (a) uses the RD-only probe and panel (b) the RI-only probe. For readability, the displayed attribution values are multiplied by 100; this rescaling is used only for visualization and does not affect the detector or the attribution computation. We select two HotpotQA responses automatically labeled as incorrect and strongly asymmetric probe confidence so that one case is dominated by RD and the other by RI. These examples are used only for qualitative interpretation, while all performance claims rely on the aggregate held-out evaluation.

Representative Cases. Panel (a) shows a dispersion-dominant entity hallucination from Qwen3-8B (HotpotQA index 301, seed 1702). For the question "Which basketball club located in Alytus, Lithuania does Anatoly Kashirov play for?", the model answers *Lietuvos Rytas*, whereas the gold answer is *BC Dzūkija*. The response contains 29 generated tokens; the frozen RD probe assigns $P(\text{wrong}) = 0.992$, compared with 0.139 from RI. Panel (b) shows an intensity-dominant formula hallucination from Mistral-7B-Instruct (HotpotQA index 15, seed 1702). For the question "Which component of shellite has the formula (ON)CHOH?", the model answers that the formula is methanol and is not a component of Shellite, whereas the gold answer is *picric acid*. The response contains 53 generated tokens; the frozen RI probe assigns $P(\text{wrong}) = 0.9997$, compared with 0.532 from RD.

E Proofs for the Theoretical Analysis of DOOR

This appendix provides the proofs for the theoretical results in Section 3.5 and clarifies their scope. All statements concern the prompt-reuse representation computed from an extracted attention block; they do not make causal claims about the origin of hallucinations.

E.1 Proof of Proposition 3.1

For one layer and attention head, let $\mathbf{P} \in \mathbb{R}_{\geq 0}^{T \times P}$ denote the generated-to-prompt attention block, with integers $T, P \geq 1$ and stabilizer $\varepsilon > 0$. Its vector of prompt-wise mean receptions is

$$\boldsymbol{\mu}(\mathbf{P}) = \frac{1}{T} \mathbf{P}^\top \mathbf{1}_T. \tag{9}$$

Let $\boldsymbol{\sigma}(\mathbf{P})$ denote the vector of prompt-wise standard deviations across generation steps. The reuse-dispersion descriptor is defined coordinate-wise as

$$c_j(\mathbf{P}) = \frac{\sigma_j(\mathbf{P})}{\max\{\mu_j(\mathbf{P}), \varepsilon\}}, \qquad j = 1, \dots, P. \tag{10}$$

Permutation invariance. Let $\widetilde{\mathbf{P}} = \mathbf{Q}_T \mathbf{P} \mathbf{Q}_P^\top$, where $\mathbf{Q}_T$ and $\mathbf{Q}_P$ are permutation matrices acting on generation steps and prompt positions, respectively. Since $\mathbf{Q}_T^\top \mathbf{1}_T = \mathbf{1}_T$,

$$\boldsymbol{\mu}(\widetilde{\mathbf{P}}) = \frac{1}{T} \widetilde{\mathbf{P}}^\top \mathbf{1}_T \tag{11}$$

$$= \frac{1}{T} \mathbf{Q}_P \mathbf{P}^\top \mathbf{Q}_T^\top \mathbf{1}_T = \mathbf{Q}_P \boldsymbol{\mu}(\mathbf{P}). \tag{12}$$

Permuting generation steps leaves each prompt-wise multiset of attention values unchanged, whereas permuting prompt positions only reorders these multisets. Therefore,

$$\boldsymbol{\sigma}(\widetilde{\mathbf{P}}) = \mathbf{Q}_P \boldsymbol{\sigma}(\mathbf{P}), \qquad \mathbf{c}(\widetilde{\mathbf{P}}) = \mathbf{Q}_P \mathbf{c}(\mathbf{P}). \tag{13}$$

The rank-profiling operator sorts each descriptor before sampling its rank curve. Consequently,

$$\Pi_K\left(\boldsymbol{\mu}(\widetilde{\mathbf{P}})\right) = \Pi_K(\boldsymbol{\mu}(\mathbf{P})), \qquad \Pi_K\left(\mathbf{c}(\widetilde{\mathbf{P}})\right) = \Pi_K(\mathbf{c}(\mathbf{P})). \tag{14}$$

Applying this argument independently to every layer and head and concatenating the resulting profiles gives Eq. (6).

Scale behavior. The two descriptors respond differently to a positive algebraic rescaling of the attention block. For $a > 0$,

$$\boldsymbol{\mu}(a\mathbf{P}) = a\boldsymbol{\mu}(\mathbf{P}), \qquad \boldsymbol{\sigma}(a\mathbf{P}) = a\boldsymbol{\sigma}(\mathbf{P}). \tag{15}$$

Positive scaling preserves the ranks of RI. Hence the RI profile scales linearly with a. For any coordinate on which the stabilizer floor is inactive both before and after scaling,

$$\frac{a\sigma_j}{a\mu_j} = \frac{\sigma_j}{\mu_j}, \tag{16}$$

so the corresponding RD value is unchanged. Thus RI retains absolute reception magnitude, whereas RD captures relative dispersion. Exact scale invariance of RD is not claimed when the denominator floor is active, because ε introduces an absolute numerical scale; RD values and their ranks may then change. □

E.2 Proof of Proposition 3.2

Fix one prompt position and suppress the layer–head and prompt-position indices. Under the marginal model $p_t = aB_t$ with $B_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\rho)$,

$$\mathbb{E}[p_t] = a\rho, \qquad \text{Var}(p_t) = a^2\rho(1-\rho). \tag{17}$$

Let $\mu^\star = \mathbb{E}[p_t] = a\rho$ and $\sigma^\star = \sqrt{\text{Var}(p_t)}$. Because $a > 0$ and $0 < \rho \leq 1$, the unstabilized population coefficient of variation $c^\star = \sigma^\star/\mu^\star$ is

$$c^\star = \frac{a\sqrt{\rho(1-\rho)}}{a\rho} = \sqrt{\frac{1-\rho}{\rho}}. \tag{18}$$

Solving $(c^\star)^2 = (1-\rho)/\rho$ gives

$$\rho = \frac{1}{1+(c^\star)^2}, \qquad a = \mu^\star\left(1+(c^\star)^2\right). \tag{19}$$

Hence the pair $(\mu^\star, c^\star)$ uniquely determines the reuse frequency ρ and active attention magnitude a. The stabilized population descriptor is $c_\varepsilon^\star = \sigma^\star/\max\{\mu^\star, \varepsilon\}$, which agrees with $c^\star$ whenever $\mu^\star \geq \varepsilon$.

Neither descriptor identifies both quantities by itself. RI alone is insufficient: for any fixed $\mu^\star \in (0,1)$ and any $\rho \in [\mu^\star, 1]$, the pair

$$(a, \rho) = \left(\frac{\mu^\star}{\rho}, \rho\right)$$

has the same mean reception $\mu^\star$. RD alone is insufficient because Eq. (18) contains no dependence on a. The paired per-token statistics $(\mu^\star, c^\star)$ therefore resolve an ambiguity that neither descriptor resolves individually under this marginal model. This result characterizes the complementarity of RI and RD at the per-token level; whether their cross-channel correspondence should be preserved after rank profiling is examined empirically by the shared-rank variant in Table 5.

Consistency of the empirical descriptors. Let $\overline{B}_T = T^{-1}\sum_{t=1}^{T} B_t$. By the strong law of large numbers, $\overline{B}_T \to \rho$ almost surely, and therefore

$$\widehat{\mu}_T = a\overline{B}_T \longrightarrow a\rho = \mu^\star \qquad \text{almost surely.} \tag{20}$$

Using the population-normalized empirical variance,

$$\widehat{\sigma}_T^2 = a^2\left(\overline{B}_T - \overline{B}_T^2\right) \longrightarrow a^2\rho(1-\rho) \qquad \text{almost surely.} \tag{21}$$

The usual $T-1$ normalization has the same asymptotic limit. For any fixed $\varepsilon > 0$, continuity gives

$$\widehat{c}_T = \frac{\widehat{\sigma}_T}{\max\{\widehat{\mu}_T, \varepsilon\}} \longrightarrow \frac{\sigma^\star}{\max\{\mu^\star, \varepsilon\}} = c_\varepsilon^\star \qquad \text{almost surely.} \tag{22}$$

If $\varepsilon < a\rho$, then $\widehat{\mu}_T > \varepsilon$ eventually almost surely, so the denominator floor becomes inactive and $\widehat{c}_T \to c^\star$ almost surely. Thus the empirical RI and RD descriptors consistently approach their population counterparts; when the population mean exceeds the stabilizer, the RD limit is the unstabilized CV used in the identification formulas above. □

Remark E.1 (Scope of the marginal model)**.** *The Bernoulli construction is not intended as a generative model of an entire row-stochastic attention matrix. It isolates one coordinate of received attention so that the distinct roles of RI and RD can be analyzed explicitly. It also does not model temporal autocorrelation: under the i.i.d. assumption, ρ represents reuse frequency rather than run-length persistence or temporal ordering.*

Remark E.2 (Random active magnitudes)**.** *The constant-amplitude assumption can be relaxed. Suppose $p_t = B_t A_t$, where $B_t \sim \mathrm{Bernoulli}(\rho)$ and the positive random magnitude A_t is independent of B_t, with mean a and active-state coefficient of variation η. Then*

$$\mathrm{CV}(p_t)^2 = \frac{1+\eta^2}{\rho} - 1. \tag{23}$$

Thus, in the more general case, the unstabilized population CV reflects both reuse frequency and variation in active attention magnitude. Proposition 3.2 uses the constant-amplitude case only to isolate the simplest setting in which the two underlying quantities are analytically identifiable; it does not assume that real attention exactly follows this model.

E.3 Proof of Theorem 3.3

We first formalize the continuous rank curve induced by a descriptor. For a nonempty vector $\mathbf{u} \in \mathbb{R}^P$, let $u_{(1)} \geq \cdots \geq u_{(P)}$ denote its descending order statistics. For $P > 1$, define the fractional rank map

$$r_P(\tau) = 1 + (P-1)\tau, \qquad \tau \in [0,1], \tag{24}$$

and define the full rank curve by linearly interpolating between adjacent order statistics indexed by $r_P(\tau)$. For $P = 1$, the rank curve is defined to be constant. This places descriptors of different lengths on the common fractional-rank domain $[0,1]$.

Information preserved by the full rank curve. Consider two descriptor vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^P$ with the same prompt length. If they differ only by a permutation, then they have the same multiset of coordinates and therefore identical descending order statistics and identical rank curves. Conversely, if their full rank curves are identical, they agree at every order-statistic knot. Their sorted vectors are

therefore identical, so the original vectors contain the same multiset of values and differ only by a permutation.

Thus, at fixed P, the full rank curve discards absolute token positions but preserves the complete multiset of descriptor values. Applying the argument independently to RI and RD also makes explicit that independent ranking discards the cross-channel correspondence between their prompt-token positions. The shared-rank ablation in Table 5 evaluates the alternative of retaining this correspondence.

Approximation by K sampled ranks. Let $q : [0,1] \to \mathbb{R}$ denote a full rank curve that is L-Lipschitz, and let $\widehat{q}_K$ be the piecewise-linear reconstruction from $K \geq 2$ uniformly spaced samples at $\tau_k = (k-1)/(K-1)$, $k = 1, \ldots, K$. Consider one grid interval $[\tau_k, \tau_{k+1}]$ of width

$$h = \frac{1}{K-1}.$$

For any $\tau = (1-\lambda)\tau_k + \lambda\tau_{k+1}$ with $\lambda \in [0,1]$,

$$\widehat{q}_K(\tau) = (1-\lambda)q(\tau_k) + \lambda q(\tau_{k+1}). \tag{25}$$

Using the triangle inequality and the Lipschitz property,

$$\begin{aligned} |q(\tau) - \widehat{q}_K(\tau)| &\leq (1-\lambda)|q(\tau) - q(\tau_k)| + \lambda|q(\tau) - q(\tau_{k+1})| \\ &\leq (1-\lambda)L\lambda h + \lambda L(1-\lambda)h \\ &= 2L\lambda(1-\lambda)h \leq \frac{L}{2(K-1)}. \end{aligned} \tag{26}$$

Taking the supremum over $\tau \in [0,1]$ gives

$$\|q - \widehat{q}_K\|_\infty \leq \frac{L}{2(K-1)}. \tag{27}$$

If a distributional score S is γ-Lipschitz with respect to the supremum norm, then

$$|S(q) - S(\widehat{q}_K)| \leq \gamma \|q - \widehat{q}_K\|_\infty \leq \frac{\gamma L}{2(K-1)}, \tag{28}$$

which proves Eq. (8). The same argument applies independently to every layer, head, and descriptor type. □

Remark E.3 (Regularity)**.** *For the piecewise-linear rank curve induced by a finite vector* $\mathbf{u}$*, one valid sample-specific Lipschitz constant is*

$$L_{\mathbf{u}} = (P-1) \max_{1 \leq i < P} \left| u_{(i)} - u_{(i+1)} \right|, \tag{29}$$

with $L_{\mathbf{u}} = 0$ *when* $P = 1$*. A uniform approximation guarantee over a collection of examples additionally requires these sample-specific constants to be uniformly bounded. The theorem concerns only the information lost when a full rank curve is represented by finitely many sampled ranks; it makes no assumption about the form or performance of the downstream classifier.*

E.4 Scope of the Guarantees

The results above establish structural properties of the DOOR representation rather than a causal theory of hallucination. Proposition 3.1 establishes invariance to permutations of generation steps and prompt positions after independent rank profiling; the accompanying scale analysis clarifies the distinct magnitude sensitivity of RI and RD. Proposition 3.2 shows, under an explicit marginal surrogate model, why the paired population mean and unstabilized CV provide complementary information about attention strength and reuse frequency. This result concerns per-token complementarity rather than identifiability of the final independently ranked representation; the effect of preserving cross-channel correspondence is evaluated empirically in Table 5. Theorem 3.3 characterizes what is retained by the full rank representation and bounds the approximation error introduced by sampling it at K ranks. Whether these representation-level properties are predictive of response errors is an empirical question addressed by the experiments in the main paper.