DISF: Detecting Hallucinations in Retrieval-Augmented Generation via Dual-path Internal State Forcing Framework

Zhe Yu[1,3,*], Wenpeng Xing[1,2,*], Wenjie Luo[2], Weize Xu[2], Lingtong Huang[2], Yourong Chen[4], Changting Lin[1,5], Meng Han[1,2,5,†]
[1]Binjiang Institute of Zhejiang University, Hangzhou, China
[2]Zhejiang University, Hangzhou, China [3]Communication University of Zhejiang, Hangzhou, China
[4]Zhejiang Shuren University, Hangzhou, China [5]GenTel.io
235703223@stu.cuz.edu.cn, {wpxing, mhan}@zju.edu.cn

Abstract

While Retrieval-Augmented Generation (RAG) grounds Large Language Models (LLMs) in external evidence, models frequently succumb to "faithfulness hallucinations" by prioritizing parametric memory over retrieved context. Existing detectors face a sharp trade-off: sampling methods are accurate but computationally prohibitive, while efficient single-pass metrics fail on confident errors due to the theoretical lack of *contrastive negative signals*. To bridge this gap, we propose **DISF**, a white-box framework that operationalizes this theoretical necessity via DUAL-PATH INTERNAL STATE-FORCING. By contrasting internal states generated with and without retrieval context, DISF captures the *Conflict*, *Drift*, and *Instability* inherent to unfaithful generation. Experiments across six backbone LLMs on two benchmarks demonstrate that DISF outperforms both unsupervised uncertainty methods and supervised internal-state baselines under a unified evaluation protocol, achieving state-of-the-art performance in hallucination detection and selective prediction.

1 Introduction

Large Language Models (LLMs) are persistently undermined by *hallucinations*—fluent but factually incorrect generations (Ji et al., 2023; Zhang et al., 2025)—alongside broader reasoning and safety vulnerabilities (Xing et al., 2025b). To mitigate this, Retrieval-Augmented Generation (RAG) grounds responses in external knowledge; however, it introduces new failure modes where models ignore or misinterpret the context, leading to "faithfulness hallucinations" (Huang et al., 2025). Consequently, robust hallucination detection remains a prerequisite for safe deployment.

Existing detection methods bifurcate into high-latency black-box sampling (Manakul et al., 2023) and efficient white-box probing (Binkowski et al., 2025). However, single-pass white-box methods face a fundamental bottleneck: Karbasi et al. (2025) prove that reliable detection is impossible using only "positive" examples, necessitating *contrastive negative signals*. This insight explains why uncertainty metrics fail on confident hallucinations—they lack the necessary comparative reference to signal errors.

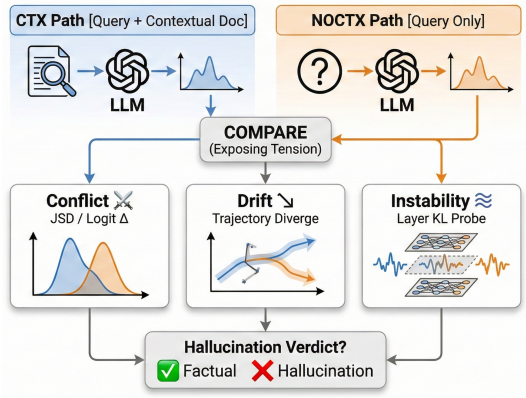

Figure 1: DISF identifies hallucinations by exposing the internal tension between context-aware (*CTX*) and context-free (*NOCTX*) inference paths.

To bridge this gap efficiently, we propose DISF, a white-box framework utilizing DUAL-PATH INTERNAL STATE-FORCING to construct the "contrastive negative signals" theoretically required for detection (Karbasi et al., 2025). By forcing the model to traverse identical response trajectories under context-rich (CTX) and context-free (NOCTX) conditions, we isolate the interplay between parametric memory and external evidence. The resulting internal divergences—quantified via *Conflict*, *Drift*, and *Instability* metrics—serve as robust indicators of hallucination, identifying when model generation decouples from retrieved context.

Our contributions are as follows:

- We propose DISF, a DUAL-PATH INTERNAL

[*]Equal contribution.
[†]Corresponding author.

STATE-FORCING framework that operationalizes the theoretical necessity of negative supervision (Karbasi et al., 2025) by quantifying the tension between parametric priors and external evidence.

- We formulate three feature families—*Conflict*, *Drift*, and *Instability*—to capture the latent distributional shifts and trajectory divergences underlying unfaithful generation, thereby enabling state-of-the-art performance on hallucination detection.

2 Related Work

2.1 Hallucination in Retrieval-Augmented Generation

While LLMs demonstrate remarkable capabilities, they are prone to *hallucinations*—generating content that is fluent but factually incorrect or unverifiable (Ji et al., 2023; Zhang et al., 2025; Huang et al., 2025). To mitigate this, RAG grounds generation in external evidence (Lewis et al., 2020; Guu et al., 2020). However, RAG introduces unique challenges, specifically "faithfulness hallucinations" where the model ignores retrieved context in favor of its parametric memory, or misinterprets the evidence (Huang et al., 2025; Niu et al., 2024). Unlike open-domain fabrication, these errors represent a failure of context adherence, necessitating specialized detection mechanisms that can disentangle parametric priors from contextual grounding (Wu et al., 2024).

2.2 Hallucination Detection Paradigms

Existing detection approaches generally fall into two categories: stochastic black-box methods and internal white-box methods.

Black-box and Sampling-based Methods. Approaches like *SelfCheckGPT* (Manakul et al., 2023) and others (Cohen et al., 2023; Mündler et al., 2023) assume hallucinated facts cause stochastic inconsistencies across sampled responses. By sampling and checking for consistency, they achieve high detection performance without model internals. However, their prohibitive computational costs (Xu et al., 2024) make them impractical for real-time RAG deployments where rapid safe defense is critical (Xing et al., 2025a).

White-box and Internal State Probing. White-box methods leverage internal signals (e.g., token probabilities, entropy, hidden states) available during a single pass (Azaria and Mitchell, 2023; Zhang et al., 2023b), analysis of which can reveal both factual anomalies and underlying safety vulnerabilities (Xing et al., 2024). Recent advances explore granular features: Chuang et al. (2024) use "Lookback Lens" on context tokens; Binkowski et al. (2025) analyze spectral properties of attention maps; and Dasgupta et al. (2025) analyze hidden-state shifts. Efficient distance-based faithfulness auditors have also emerged for residual streams (Yu et al., 2026b). Supervised internal-state methods also show promise: *FactoScope* (He et al., 2024) decodes hidden states via projection, *SpikeScore* (Liang et al., 2024) utilizes cross-layer entropy spikes, and UQ heads (Kossen et al., 2024) probe attention patterns. While effective, they often struggle with "confident hallucinations," where internal confidence remains high despite factual errors (Simhi et al., 2025).

2.3 Theoretical Foundations and Contrastive Dynamics

The limitation of single-pass white-box methods can be explained by recent theoretical findings. Karbasi et al. (2025) formally proved that automated hallucination detection is fundamentally impossible if the detector is trained solely on "positive" examples (i.e., the generated text alone). They argue that reliable detection necessitates *contrastive negative signals* or explicit feedback to distinguish between plausible text and hallucinations.

This theoretical insight aligns with mitigation strategies that employ contrastive dynamics, such as *Context-Aware Decoding (CAD)* (Shi et al., 2024), which amplifies the difference between output probabilities with and without context to improve faithfulness, and *FRANQ* (Andriushchenko et al., 2025), which decomposes branch-wise scalar uncertainties from CTX/NOCTX runs. Additionally, the exploration of explicit CTX/NOCTX representation shifts has begun showing promise for domain-specific risk triage, such as combating hallucinations and evidence gaps in medical LLMs (Yu et al., 2026a). However, CAD targets generation rather than detection, and FRANQ operates at the output-level uncertainty decomposition.

Our work bridges this gap by operationalizing the theoretical requirement for negative signals (Karbasi et al., 2025) within a detection framework. Unlike standard white-box probes that analyze a single generation trace (Azaria and Mitchell, 2023; Binkowski et al., 2025) or output-level CTX/NOCTX divergences (Andriushchenko et al.,

2025), DISF explicitly constructs a dynamic reference via DUAL-PATH INTERNAL STATE-FORCING. By contrasting the *internal trajectory dynamics* of a context-rich path against a context-free path across model depth, we expose the "tug-of-war" between parametric memory and retrieved evidence, thereby providing a robust indicator of faithfulness hallucinations.

3 Methodology

We propose DISF, a white-box RAG hallucination detection framework requiring no additional sampling or labeled negative data. Following the problem formalization (Section 3.1) and contrastive theoretical foundation (Section 3.2), we detail three token-level feature families: knowledge conflict, trajectory drift, and internal instability (Section 3.3). Finally, we describe their sparse Top-K aggregation and classification via a lightweight XGBoost model (Section 3.4).

3.1 Problem Formulation

DISF targets the problem of *Response-Level Hallucination Detection* in RAG. Formally, given a user query q, a retrieved document set $\mathcal{D}$, and a model-generated response $r = \{y_1, y_2, \ldots, y_T\}$, our objective is to learn a binary decision function $f(r, q, \mathcal{D}) \rightarrow \{0, 1\}$, where 1 indicates the presence of hallucinatory content (either factual fabrication or unfaithfulness to $\mathcal{D}$).

3.2 Theoretical Framework

Existing hallucination detection methods fall into high-latency black-box sampling (Manakul et al., 2023) and white-box *internal state probing (ISP)* (Azaria and Mitchell, 2023). We reconceptualize RAG hallucinations as *knowledge conflict*: dissonance between the model's parametric memory and retrieved non-parametric evidence.

Crucially, DISF is formally grounded as a *controlled counterfactual intervention*. For a given response token sequence r, we enforce identical generation trajectories under two conditions: the factual contextual path (CTX: $[q; \mathcal{D}]$) and the counterfactual parametric-only path (NOCTX: $[q]$). By aligning both branches on the identical token prefix via teacher-forcing, the resulting discrepancy precisely isolates context-driven internal effects from sampling-path variance, ensuring that all observed internal state deviations are attributable to knowledge conflict rather than stochastic generation artifacts. Under local linearization, internal-state discrepancies directly induce prediction discrepancies, motivating the three feature families targeting distinct etiological roots. Furthermore, DISF establishes a strict *in-policy* methodology: during deployment, the CTX features are evaluated on responses directly generated by a target backbone, while the NOCTX features are extracted from the identical backbone via counterfactual replay, eliminating confounding factors from cross-model representation shifts.

3.3 Heterogeneous Feature Engineering

Building on the dual-path contrastive reference, we extract three families of token-level features $\mathbf{x}_t \in \mathbb{R}^d$ at each generation step t. These features target distinct etiological roots of RAG hallucinations: (1) **conflict features** (Section 3.3.1), quantifying the tension between retrieved evidence and parametric memory; (2) **drift features** (Section 3.3.2), capturing semantic trajectory divergence across layers; and (3) **instability features** (Section 3.3.3), measuring internal uncertainty and indecision. The following subsections detail each family.

3.3.1 Conflict Features: Evidence vs. Memory

Hallucination in RAG is frequently precipitated by knowledge conflict—a phenomenon where the model's intrinsic parametric priors contradict the retrieved external evidence. This dynamic, often characterized as a "tug-of-war" between memory and context (Wu et al., 2024), manifests when the model over-relies on stubborn internal beliefs, thereby suppressing valid retrieved information.

To quantify this dissonance, we measure the divergence between the predictive distributions of the CTX and NOCTX decoding trajectories. We introduce the Logit Shift, $S_{logit}(t)$, to capture the marginal contribution of the retrieved context $\mathcal{D}$ to the target token's likelihood:

$$S_{logit}(t) = \text{Logit}(y_t \mid [q; \mathcal{D}]) - \text{Logit}(y_t \mid [q]) \tag{1}$$

A significant positive shift indicates that the external evidence $\mathcal{D}$ acts as a constructive signal, boosting the generation probability of y_t. Crucially, a *negative* shift signals that the presence of context actively suppresses the token's likelihood. This suppression serves as a strong proxy for hallucination, implying that the model is generating y_t based on parametric priors that directly contradict the provided evidence.

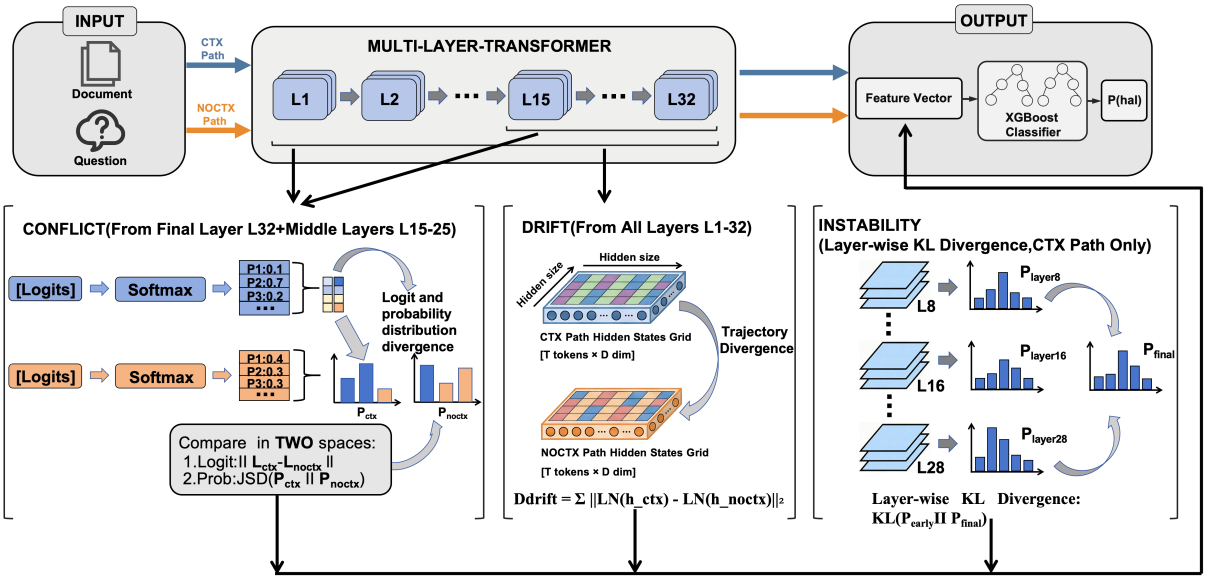

Figure 2: DISF framework overview. The process operates in three stages: (1) Heterogeneous Feature Engineering: Leveraging a dual-path internal state-forcing strategy (CTX vs. NOCTX) to extract token-level internal dynamics, categorized into *Conflict*, *Drift*, and *Instability* features; (2) Sparse Aggregation: Aggregating token-level signals into a fixed-length response vector using statistical operators; and (3) Inference: Utilizing a supervised classifier to predict the probability of hallucination at the response level.

Complementing this local metric, we compute the Jensen-Shannon Divergence (JSD) to quantify the global distributional perturbation. A high JSD between these distributions indicates that the retrieval context has significantly altered the model's cognitive state, marking potential areas of conflict or correction.

3.3.2 Drift Features: Trajectory Divergence

While conflict features capture the static disagreement between memory and evidence, hallucinations also manifest as dynamic anomalies in the model's reasoning process. Building on Chuang et al. (2023)'s finding that factual knowledge is stratified and evolves across transformer layers, we propose analyzing the Layer-wise Semantic Dynamics (LSD). We conceptualize the sequence of hidden states across layers $l \in \{1, \ldots, L\}$ as a continuous trajectory in the semantic space.

We quantify the stability of this reasoning chain by computing the Trajectory Divergence (D_{drift}) between the *Contextual* and *Parametric* paths. This is formulated as the cumulative Euclidean distance between normalized hidden states:

$$D_{\text{drift}}(t) = \sum_{l=1}^{L} \left\| (h_{t,l}^{\text{ctx}}) - (h_{t,l}^{\text{noctx}}) \right\|_2 \qquad (2)$$

A sudden spike in D_{drift} indicates a *distribution shift* in the internal activation space. This suggests that without external evidence, the model is forced to radically alter its activation path to maintain the generation of token y_t, implying that the generated content lacks robust parametric grounding and is highly dependent on the (potentially misinterpreted) retrieval context.

3.3.3 Instability Features: Internal Consistency Quantification

While final-layer token probability is a conventional proxy for uncertainty, LLMs frequently exhibit miscalibration, often assigning high confidence to hallucinated content due to the "likelihood trap" (Manakul et al., 2023). To capture uncertainty manifesting earlier in the generation process, we adopt a depth-wise probing approach leveraging the Logit Lens technique.

We hypothesize that truthful generation maintains semantic stability across layers, whereas hallucinations trigger internal decoherence. We quantify this by projecting the hidden state $h_{t,\ell}$ of an intermediate layer ℓ into the vocabulary space to obtain an early predictive distribution $P_\ell(y_t)$. We then compute the Layer-wise Kullback-Leibler Divergence between this intermediate distribution and

the final output distribution $P_L(y_t)$:

$$I_{\text{instable}}(t) = \text{KL}\left(P_\ell(y_t) \,\|\, P_L(y_t)\right) = \sum_{v \in \mathcal{V}} P_\ell(v) \log \frac{P_\ell(v)}{P_L(v)} \quad (3)$$

where ℓ represents an early-to-mid layer index (e.g., $L/2$). A high divergence $I_{instable}$ indicates a state of *internal dissonance*, where the model's lower-level factual encoding contradicts its final semantic projection. This aligns with findings by Chuang et al. (2023) and Azaria and Mitchell (2023), who suggest that truthfulness is often stratified across layers and that significant shifts in the predictive distribution across the network depth are precursors to factual fabrication.

3.4 Sparse Aggregation and Inference

The feature extraction process described in Section 3.3 yields a variable-length sequence of token-level indicators $\mathbf{X} \in \mathbb{R}^{T \times D}$. A critical challenge in response-level detection is that hallucinations are often *sparse* and *localized*—manifesting as specific fabricated entities or relation errors within a largely fluent text (Niu et al., 2025). Consequently, naive global pooling strategies (e.g., averaging over the entire sequence) risk diluting these salient anomaly signals with the noise from factual tokens.

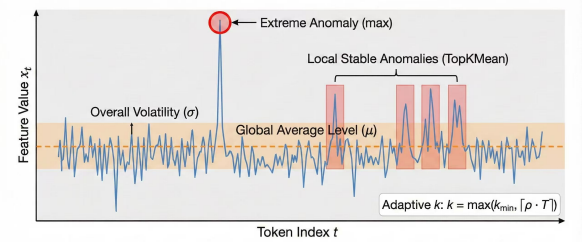

Figure 3: Illustration of DISF pooling operators on a token-level feature sequence.

To address this, we propose the Sparse Aggregation framework with two stages. First, we selectively distill the most discriminative signals (peaks of conflict or drift; see Figure 3) into a fixed-dimensional vector using Top-K Mean Aggregation (Section 3.4.1). Second, this dense representation is fed to a lightweight classifier (Section 3.4) to output the hallucination probability, achieving a balance between detection sensitivity and computational efficiency.

3.4.1 Top-K Mean Aggregation

To operationalize the sparse detection principle, we treat the sequence of token-level features as a heavy-tailed distribution, where valid tokens constitute the low-signal background and hallucinations manifest as outliers in the upper tail. We employ an adaptive ranking mechanism to isolate these outliers.

First, we define a dynamic filtering window K proportional to the response length T, ensuring scale invariance across short and long generations:

$$K = \max(k_{min}, \lceil \rho \cdot T \rceil) \quad (4)$$

where $\rho \in (0, 1]$ is a sensitivity hyperparameter (e.g., $\rho = 0.1$ targets the top 10% suspicious tokens).

Crucially, we perform the aggregation *independently* for each feature dimension d. This design choice acknowledges that different indicators are not necessarily synchronous; for instance, a *Conflict* signal (S_{logit}) may peak at a named entity, while an *Instability* signal ($I_{instable}$) may surge at a relational verb. The aggregated vector $\mathbf{v} \in \mathbb{R}^D$ is computed as:

$$v_d = \frac{1}{K} \sum_{t \in \mathcal{T}_d} \mathbf{X}_{t,d}, \quad \text{s.t. } \mathcal{T}_d = \underset{t}{\text{argtop-}K}(\mathbf{X}_{:,d}) \quad (5)$$

By decoupling the selection indices $\mathcal{T}_d$ across dimensions, our method functions as a *Multi-Channel Soft Max-Pooling*, effectively capturing the union of all salient anomalies regardless of their temporal alignment (Niu et al., 2025).

3.4.2 Classifier Design

We employ a gradient boosting framework (XGBoost) rather than deep neural networks (DNNs) driven by two constraints. First, Feature Heterogeneity: Our inputs combine bounded probabilities (e.g., P_{conf}) with unbounded divergences (e.g., D_{KL}). Tree ensembles inherently handle such tabular heterogeneity without the complex normalization required by DNNs. Second, Data Scarcity: High-quality hallucination datasets are notably scarce and imbalanced (Tang et al., 2024; Alansari and Luqman, 2025). As theoretically established by Karbasi et al. (2025), reliable detection necessitates expert-labeled negative examples, which are costly to curate. In this low-resource regime ($N < 10^4$), XGBoost offers superior generalization over over-parameterized MLPs.

Training and Inference Strategy. To mitigate overfitting and source-specific biases, we employ a GroupKFold cross-validation strategy, grouping samples by their retrieval source ID. This ensures

that response variations from the same document do not leak between folds. Unlike standard classification which defaults to $\tau = 0.5$, we dynamically calibrate the decision threshold τ to address the class imbalance inherent in hallucination datasets. Formally, let $\hat{\mathbf{y}}_{\text{oof}}$ denote the out-of-fold probability predictions collected during cross-validation. The optimal threshold τ^* is determined by maximizing the F1-score:

$$\tau^* = \underset{\tau \in [1]}{\operatorname{argmax}} \, \mathrm{F1}\left(\mathbf{y}_{\text{train}}, \, \mathbb{I}(\hat{\mathbf{y}}_{\text{oof}} > \tau)\right) \quad (6)$$

where $\mathbb{I}(\cdot)$ is the indicator function. During inference, the computational overhead is $O(1)$, which is negligible compared to the $O(T^2)$ decoding cost or the sampling overhead of black-box baselines like *SelfCheckGPT* (Manakul et al., 2023).

4 Experiments

Our experiments aim to validate our mechanistic framework by addressing the following three research questions:

RQ1 (Effectiveness): Does expert-labeled supervision significantly outperform zero-shot consistency, validating the necessity of contrastive signals for hallucination detection?

RQ2 (Dynamics): Can internal signals—*Conflict*, *Drift*, and *Instability*—reliably predict hallucination severity during long-form generation?

RQ3 (Efficiency): What are the computational overhead and scaling properties of DISF for practical, long-sequence deployment?

4.1 Experiment Setup

Datasets. We evaluate on two complementary benchmarks (detailed statistics in Appendix B): (1) *RagTruth* (Niu et al., 2024), a large-scale multi-task RAG corpus spanning QA and summarization. We use the official split ($\sim$2,500 train / $\sim$450 test samples per model) and employ 5-fold *GroupKFold* cross-validation grouped by `source_id` to ensure document-level isolation. (2) *HalluRAG* (Ridder and Schilling, 2024), a knowledge-conflict benchmark where each prompt yields two competing answers. We apply strict group-level splitting (70%/15%/15%) to prevent prompt-level leakage, using the *answerable* subset to stress-test context prioritization.

Baselines. We compare DISF against twelve baselines spanning five detection paradigms (detailed descriptions in Appendix E):

(1) Uncertainty-based methods estimate hallucination risk via token-level probabilities or entropy: *Perplexity* (Ren et al., 2023), *LN-Entropy* (Malinin and Gales, 2020), and *Focus* (Zhang et al., 2023a).

(2) Sampling-based methods measure consistency across multiple stochastic responses: *SelfCheckGPT* (Manakul et al., 2023) and *EigenScore* (Chen et al., 2024).

(3) Verbalization methods prompt LLMs to self-evaluate or verify claims: *P(True)* (Kadavath et al., 2022) and *RefChecker* (Hu et al., 2024).

(4) Interpretability-based methods leverage mechanistic signals to decouple context and parametric knowledge: *ReDeEP* (Sun et al., 2024).

(5) Supervised internal-state methods train classifiers on internal representations: *FactoScope* (He et al., 2024), *UQ Heads* (Kossen et al., 2024), *LookbackLens* (Chuang et al., 2024), and *SpikeScore* (Liang et al., 2024). All are evaluated under the same response-level protocol (same split, readout, and metrics; see Appendix F).

Metrics. We report AUROC and AUPRC for ranking quality, F1 at optimal thresholds for binary detection, and PCC to quantify alignment with hallucination severity. We additionally report deployment-oriented selective prediction metrics—*Prediction-Rejection Ratio (PRR)*, *AURC*, and *Risk@Coverage*—in Appendix G.4. Complete protocol details are in Appendix D.

Backbone LLMs. We test on six open-weight models: `LLaMA2-7B-Chat`, `LLaMA2-13B-Chat`, `LLaMA3-8B-Instruct`, `Mistral-7B-Instruct-v0.1`, `Qwen3-8B`, and `Qwen3-14B`, with the detector accessing each model's hidden states directly. The first four backbones constitute the main evaluation; the Qwen3 models validate generalization to newer-generation architectures (Appendix G.1).

4.2 Experiment Results

4.2.1 Effectiveness (RQ1)

Table 1 summarizes response-level detection performance across four model backbones against unsupervised and zero-shot baselines. Our analysis highlights the following key observations.

Cross-Dataset Generalization. The consistent superiority of DISF across diverse benchmarks demonstrates strong cross-dataset generalization. Despite differences in dataset construction and hallucination types (*RagTruth* for general tasks

Table 1: Response-level hallucination detection on *RAGTruth* and *HalluRAG* test sets. We report AUROC, AUPRC, and PCC. All methods use the same response-level protocol. Missing values are denoted by "-".

Method	RAGTruth			HalluRAG		
	AUROC↑	AUPRC↑	PCC↑	AUROC↑	AUPRC↑	PCC↑
LLaMA2-7B						
Perplexity	0.5307	0.6107	-0.0644	0.5124	0.5210	-0.0123
LN-Entropy	0.6702	0.7370	0.2146	0.6432	0.6120	0.1842
Focus	0.6144	0.6782	0.1025	0.6486	0.6973	0.2541
SelfCheckGPT	0.5970	0.5734	0.0327	0.6066	0.4787	0.0147
EigenScore	0.5867	0.5527	-0.0685	0.8238	0.8031	0.5276
P(True)	0.6509	0.7398	0.1990	0.7484	0.7212	0.3625
RefChecker	0.5951	0.6189	0.0260	0.5789	0.4455	-0.0876
ReDeEP	0.7522	0.7528	0.1036	0.9248	0.9266	0.5059
DISF	**0.7947**	**0.8133**	**0.2043**	**0.9582**	**0.9573**	**0.6880**
LLaMA2-13B						
Perplexity	0.4230	0.4885	-0.1823	0.8462	0.8719	0.2553
LN-Entropy	0.7146	0.7160	0.3525	0.6397	0.7135	0.2627
Focus	0.5902	0.5027	0.1032	0.6592	0.7228	0.2662
SelfCheckGPT	0.6248	0.6123	-0.0076	0.6392	0.6148	0.2515
EigenScore	0.7929	0.7793	0.1692	0.7148	0.7825	0.3424
P(True)	0.5413	0.4900	0.0410	0.7912	0.7877	0.5469
RefChecker	0.7774	0.7770	0.1497	0.7659	0.7802	0.3221
ReDeEP	0.8311	0.8363	0.1959	0.8326	0.8452	0.4481
DISF	**0.8578**	**0.8558**	**0.2281**	**0.8655**	**0.9027**	**0.5067**
LLaMA3-8B						
Perplexity	0.8189	0.7003	0.0350	0.8214	0.8342	0.1523
LN-Entropy	0.8076	0.6843	0.4581	0.7915	0.7841	0.3842
Focus	0.7326	0.4625	0.2890	0.5442	0.8032	-0.0379
SelfCheckGPT	0.6039	0.6010	-0.0431	0.5964	0.5979	0.1923
EigenScore	0.8049	0.7286	0.2676	0.6708	0.3845	0.2777
P(True)	0.7415	0.5832	0.3690	0.7399	0.8070	0.2094
RefChecker	0.8146	0.7909	0.3295	0.6899	0.7081	0.4212
ReDeEP	0.8677	0.8496	0.3685	0.8331	0.8111	0.4411
DISF	**0.9354**	**0.8849**	**0.3849**	**0.8489**	**0.8410**	**0.5840**
Mistral-7B						
Perplexity	0.5995	0.6891	0.0517	0.5841	0.5923	0.0421
LN-Entropy	0.7462	0.8024	0.3158	0.7215	0.7341	0.2910
Focus	0.6988	0.6557	0.2420	0.6676	0.7245	0.2874
SelfCheckGPT	0.6474	0.6288	0.0713	0.7332	0.6285	0.4896
EigenScore	0.7378	0.7510	0.0594	0.6200	0.6532	0.1984
P(True)	0.4241	0.4238	-0.0387	0.8720	0.8833	0.5666
RefChecker	0.6576	0.6748	0.1512	0.7448	0.7512	0.1255
ReDeEP	0.8271	0.8357	0.1898	0.8871	0.8954	0.5892
DISF	**0.8448**	**0.8641**	**0.2162**	**0.9246**	**0.9416**	**0.6681**

and *HalluRAG* for knowledge conflicts), DISF robustly outperforms all baselines. For instance, on `LLaMA2-7B`, it achieves 0.795 AUROC on *RagTruth* and 0.958 on *HalluRAG*, maintaining a significant margin over the strongest baselines in both datasets.

Comparison with Uncertainty and Zero-Shot Methods. Classic uncertainty baselines—*Perplexity* and *LN-Entropy*—exhibit fundamental limitations in RAG scenarios. Despite being computationally efficient, Perplexity achieves only 0.423–0.819 AUROC across models on *RagTruth*, with notably weak or even negative PCC (e.g., -0.182 on `LLaMA2-13B`), indicating that token-level confidence alone fails to capture hallucination severity. LN-Entropy improves upon Perplexity by incorporating sequence-level normalization, yet remains inferior to mechanistic approaches.

Table 2: Comparison with supervised internal-state baselines (8-setting average across 4 backbones $\times$ 2 datasets). All methods use the same response-level protocol. Full per-backbone tables in Appendix F.

Method	AUROC↑	AUPRC↑	PRR↑
LookbackLens	0.7068	0.7269	0.4787
UQ Heads	0.7973	0.8231	0.5907
SpikeScore	0.7776	0.7722	—
FactoScope	0.8779	0.8826	0.7758
DISF (Ours)	**0.8891**	**0.8986**	**0.7830**

Our supervised DISF framework also substantially outperforms unsupervised zero-shot baselines. *SelfCheckGPT*, relying on stochastic sampling and NLI-based consistency without labeled data, achieves only 0.597–0.647 AUROC on *RagTruth*, lagging DISF by 20–30 points. *EigenScore*, despite using internal states, yields 0.587–0.805 AUROC via unsupervised eigenvalue decomposition. These results empirically validate Karbasi et al. (2025), confirming that surface-level uncertainty signals cannot disentangle parametric priors from contextual grounding, and underscoring the value of mechanistically-aware supervision.

Comparison with Supervised Internal-State Methods. Table 2 compares DISF against four supervised baselines under the same response-level protocol (identical split, GroupKFold strategy, and XGBoost readout). DISF achieves the highest 8-setting mean across all primary metrics. Notably, *FactoScope*—which also leverages Logit Lens projections—achieves competitive AUROC (0.8779) but falls short on AUPRC and PRR, confirming that trajectory-level state dynamics (Drift) contribute discriminative power beyond output-layer features. The gap widens for methods relying on single-signal families: *LookbackLens* (attention-only, 0.7068 AUROC) and *UQ Heads* (0.7973 AUROC). Across individual settings, DISF wins AUROC on 6/8, AUPRC on 6/8, and Risk@80%Cov on 7/8 settings. We also verify that replacing XGBoost with a simpler logistic regression readout preserves DISF's advantage (Appendix F), confirming that the gains originate from feature quality rather than downstream classifier capacity.

Generalization to Newer Backbones. To validate robustness on newer model generations, we additionally evaluate DISF on `Qwen3-8B` and

Table 3: Ablation study of key dynamic features on *RagTruth* test set. **Bold**/underline denote best/second-best per column.

		LLaMA2-7B		LLaMA2-13B		LLaMA3-8B	
Variant	**#Feat**	AUROC↑	PCC↑	AUROC↑	PCC↑	AUROC↑	PCC↑
Ours (Full)	145	**0.7947**	0.2043	**0.8578**	0.2281	**0.9354**	0.3849
w/o Conflict	93	0.7821	0.1899	0.8507	**0.2284**	0.9341	0.3887
w/o Drift	85	0.7729	**0.2202**	0.8539	0.1931	0.9047	0.3712
w/o Instability	113	0.7878	0.1488	0.8177	0.1763	0.9061	0.3491
Only Conflict	53	0.7210	0.0646	0.8217	0.1787	0.9313	**0.3889**
Only Drift	61	0.7583	0.1321	0.8421	0.2021	0.9341	0.3728
Only Instability	33	0.7186	0.1765	0.8317	0.1664	0.8943	0.3220

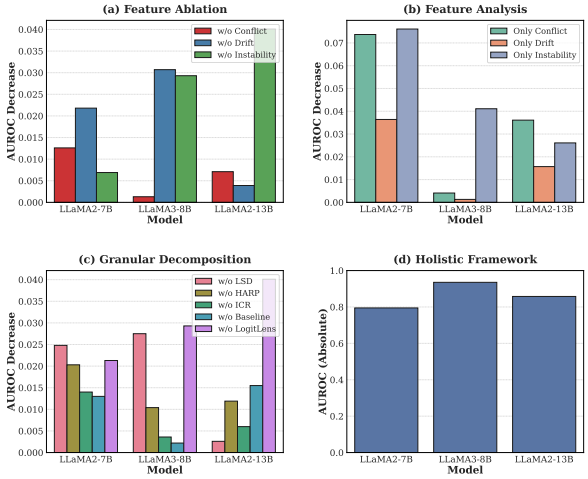

Figure 4: Ablation analysis of conflict, drift, and instability signals (a–b), illustrating the role of granular components (c) in sustaining holistic performance across model scales (d).

Qwen3-14B under the same protocol. DISF achieves strong performance (e.g., Qwen3-14B on HalluRAG: 0.9146 AUROC), demonstrating that the proposed signals remain informative on newer, stronger architectures. Leave-one-model-out (LOMO) cross-model transfer further shows meaningful zero-shot generalization (overall mean 0.7959/0.7271 AUROC/AUPRC; Appendix G.2).

Correlation with Hallucination Severity. Beyond binary detection, DISF demonstrates stronger alignment with hallucination severity as measured by PCC. On *RagTruth*, DISF achieves 0.204–0.385 PCC, outperforming all baselines. The improvement is most evident on LLaMA3-8B (+1.6 points over *ReDeEP*) and LLaMA2-7B (+10.1 points). This indicates that our dynamic signals not only discriminate hallucinated from faithful responses, but also capture the degree of factual deviation.

4.2.2 Dynamic Signal (RQ2)

To answer **RQ2**, we conduct comprehensive ablation experiments (Appendix Table 3, Table 11) of *Conflict* (evidence-memory suppression via negative Logit Shifts or high JSD), *Drift* (layer-wise trajectory divergences), and *Instability* (internal KL-based decoherence). As depicted in Figure 4(d), LLaMA3-8B establishes a superior baseline, followed by LLaMA2-13B and LLaMA2-7B. A key insight emerges: hallucination "fingerprints" are *architecture-dependent* rather than scale-dependent. The LLaMA2→LLaMA3 architectural shift produces qualitatively different detection signatures, an effect far exceeding the modest gains from scaling within the same family (7B→13B).

Model-Specific Feature Dependencies. Ablation results (Figure 4(a)) reveal architecture-driven rather than scale-dependent feature sensitivities. LLaMA2-13B exhibits a *bottleneck dependency* on *Instability* features: removing inter-layer KL divergence triggers the largest AUROC drop, suggesting hallucination detection hinges on internal prediction fluctuations rather than semantic consistency. In contrast, LLaMA3-8B shows remarkable resilience to *Conflict* removal, instead relying on *Drift* features for trajectory-based detection—reflecting a smoother hierarchical evolution that prioritizes semantic stability. Notably, *Drift* serves as a *universal core signal*: both LLaMA2-7B and LLaMA3-8B are highly sensitive to its removal, while LLaMA2-13B uniquely prioritizes internal instability over trajectory shifts.

Single-Feature Sufficiency. The "Only" configuration analysis (Figure 4(b)) reveals divergent patterns: LLaMA2-7B suffers steep degradation across all single-feature settings, necessitating *full heterogeneous fusion*. Conversely, LLaMA3-8B maintains robust performance under *Only Conflict* or *Only Drift*, demonstrating that either signal alone provides sufficient discriminative power. For LLaMA3, *Instability* acts as a secondary signal prone to miscalibration when used in isolation.

Granular Component Analysis. Fig. 4(c) decomposes signal families into specific operators, revealing two critical trends in the internal heuristics of hallucination detection. **(1) Centrality of LogitLens and LSD:** *LogitLens* (inter-layer KL divergence) and *LSD* (trajectory dynamics) emerge as the dual pillars of performance. **(2) Fragility of LLaMA2-7B:** The 7B model is the most fragile variant, lacking the feature redundancy found in its

larger or more advanced counterparts.

Summary. LLaMA models exhibit distinct feature dependencies driven by architectural evolution. `LLaMA3-8B` achieves peak performance (AUROC 0.9354) via a semantic-centric mechanism where *Conflict* or *Drift* possess "independent sufficiency." `LLaMA2-13B` shows bottleneck dependency on *Instability*, while `LLaMA2-7B` requires full feature fusion. Overall, *Drift* serves as a universal core signal across architectures.

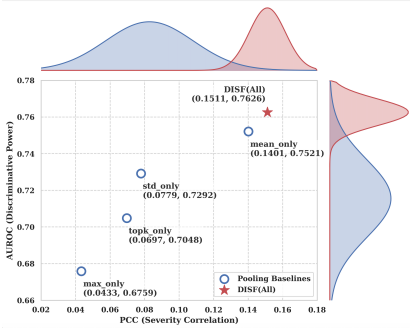

Figure 5: **Effectiveness of Pooling Strategies.** Comparison of different aggregation operators on `LLaMA2-7B`.

Effectiveness of Pooling Strategies. Analysis of different aggregation operators in Figure 5 reveals that signal aggregation is critical for detection reliability. *Max-only* pooling performs poorest (AUROC 0.6759), as reliance on single-point extremes makes the detector hyper-sensitive to transient noise. In contrast, *mean-only* pooling significantly improves performance (AUROC 0.7521), confirming that hallucinations are *span-level phenomena*: the model generates coherent but incorrect multi-token spans rather than isolated anomalies. Our *Full* strategy achieves state-of-the-art AUROC of 0.7947 by integrating *TopK* (focusing on the riskiest token subset) with *Mean/Std* for global calibration.

4.2.3 Efficiency (RQ3)

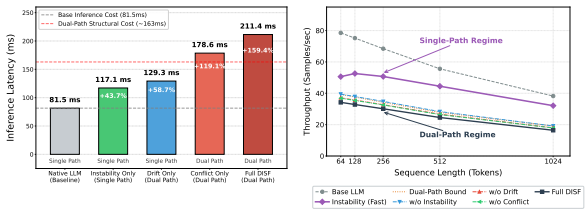

Figure 6: Latency analysis of DISF on `LLaMA2-7B`.

The DISF framework exhibits high algorithmic lightweightness through a dual-regime architecture (Figure 6).

Single-Path Regime. The *Instability-only* configuration adds only 43.7% overhead (117.1ms vs. 81.5ms baseline), retaining approximately 85% of native LLM throughput. This mode functions as a real-time filter suitable for latency-sensitive applications.

Dual-Path Regime. The full DISF configuration operates at 211.4ms (+159.4% overhead), but this cost is dominated by the structural requirement of running two forward passes (CTX and NOCTX), not by feature extraction complexity. Ablating individual feature modules yields nearly identical throughput, confirming that Drift, Conflict, and Instability computations are negligible compared to the dual forward-pass overhead.

Long-Context Amortization. DISF demonstrates superior scalability in RAG scenarios: its $\mathcal{O}(n)$ feature computation allows relative overhead to decrease as sequence length increases. To further reduce deployment cost, we propose three practical recipes—cascade filtering, CTX cache reuse, and batched NOCTX replay—detailed in Appendix H.

5 Conclusion

This paper introduced DISF, a mechanistic framework that leverages a *dual-path internal state-forcing* strategy to detect hallucinations in RAG. Unlike existing heuristics, DISF treats hallucinations as dynamic internal phenomena, capturing the latent tension between parametric priors and retrieved evidence through: *conflict*, *drift*, and *instability* features. Our work yields two primary insights. Theoretically, we demonstrate that while internal states contain rich diagnostic signals, expert-labeled negative supervision is essential to resolve their inherent ambiguity. Empirically, DISF achieves state-of-the-art performance across six backbone LLMs, outperforming both unsupervised models and structured supervised baselines under a unified evaluation protocol. Extensive experiments further confirm its strong generalization capabilities across newer architectures and cross-model transfer scenarios. To mitigate overhead, we also introduce flexible deployment recipes (e.g., cascade and batched modes) that effectively balance quality and efficiency. Future work will extend these principles to black-box and API-only environments via probabilistic approximation, advancing the development of self-calibrating LLM guardrails.

Limitations

Despite the state-of-the-art performance achieved by DISF in detecting hallucinations within Retrieval-Augmented Generation (RAG) systems, we acknowledge several limitations inherent to our framework:

Dependency on White-box Access (API-Only Incompatibility). DISF fundamentally relies on accessing the model's internal dynamics, including hidden states across layers and output logit distributions, to compute the *Conflict*, *Drift*, and *Instability* feature families. While this white-box approach captures nuanced uncertainty signals that black-box methods miss, it inherently restricts the applicability of DISF strictly to open-weight models where full structural access is permitted. Consequently, direct black-box compatibility with proprietary API-only models (e.g., GPT-4, Claude 3) represents an explicit scope boundary for full DISF, as such systems do not expose the required internal trajectory data.

Potential Risks

While DISF demonstrates robust performance in detecting RAG hallucinations, we identify three critical risks associated with it:

The "Confident Hallucination" Paradox. DISF relies heavily on *Drift* and *Instability* features, which assume that hallucinations manifest as internal tension between parametric priors and retrieved evidence, or as distributional uncertainty. A significant risk arises with models that have undergone extensive Reinforcement Learning from Human Feedback (RLHF) to maximize coherence. Such models may exhibit "sycophantic" behavior, generating hallucinations with high internal confidence and low entropy, effectively suppressing the *Drift* and *Instability* signals DISF relies upon. In these cases, the model is "confidently wrong," potentially bypassing the detector.

Acknowledgements

This research was supported by the "Open Competition" Project of Hangzhou High-tech Zone (Binjiang) (Grant No. 2025JBGS-PT004).

References

Aisha Alansari and Hamzah Luqman. 2025. Large language models hallucination: A comprehensive survey. *arXiv preprint arXiv:2510.06265*.

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2025. Franq: Faithful hallucination detection via feature-guided rag with no questions asked. *arXiv preprint arXiv:2505.21072*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. 2025. Hallucination detection in llms using spectral features of attention maps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24365–24396.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. *arXiv preprint arXiv:2407.07071*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*.

Sharanya Dasgupta, Sujoy Nath, Arkaprabha Basu, Pourya Shamsolmoali, and Swagatam Das. 2025. Hallushift: Measuring distribution shifts towards hallucination detection in llms. *arXiv preprint arXiv:2504.09482*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Jinwen He, Yujia Zhan, Zekun Qiu, Yichi Deng, Wenxuan Wang, Jiahao Lin, and Lidong Bing. 2024. Llm factoscope: Uncovering llms' factual discernment through measuring inner states. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models. *arXiv preprint arXiv:2405.14486*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

ACM Transactions on Information Systems, 43(2):1–55.

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

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

Amin Karbasi, Omar Montasser, John Sous, and Grigoris Velegkas. 2025. (im) possibility of automated hallucination detection in large language models. *arXiv preprint arXiv:2504.17004*.

Jannik Kossen, Yarin Gal, and Tom Rainforth. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chen Liang, Zhijiang Tian, and Qi Zhang. 2024. Beyond in-domain detection: Spikescore for cross-domain hallucination detection. *arXiv preprint arXiv:2505.08200*.

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

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878.

Mengjia Niu, Hamed Haddadi, and Guansong Pang. 2025. Robust hallucination detection in llms via adaptive token selection. *arXiv preprint arXiv:2504.07863*.

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

Fabian Ridder and Malte Schilling. 2024. The hallurag dataset: Detecting closed-domain hallucinations in rag applications using an llm's internal states. *arXiv preprint arXiv:2412.17056*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.

Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. 2025. Trust me, i'm wrong: High-certainty hallucinations in llms. *arXiv preprint arXiv:2502.12964*.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, and Han Li. 2024. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, et al. 2024. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4455–4480.

Kevin Wu, Eric Wu, and James Zou. 2024. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. *Advances in neural information processing systems*, 37:33402–33422.

Wenpeng Xing, Mohan Li, Chunqiang Hu, Haitao Xu, Ningyu Zhang, Bo Lin, and Meng Han. 2024. Latent fusion jailbreak: Blending harmful and harmless representations to elicit unsafe llm outputs. *arXiv preprint arXiv:2508.10029*.

Wenpeng Xing, Zhonghao Qi, Yupeng Qin, Yilin Li, Caini Chang, Jiahui Yu, Changting Lin, Zhenzhen Xie, and Meng Han. 2025a. Mcp-guard: A defense framework for model context protocol integrity in large language model applications. *arXiv preprint arXiv:2508.10991*.

Wenpeng Xing, Lanyi Wei, Haixiao Hu, Jingyi Yu, Rongchang Li, Mohan Li, Changting Lin, and Meng Han. 2025b. Sproutbench: A benchmark for safe and ethical large language models for youth. *arXiv preprint arXiv:2508.11009*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Zhe Yu, Wenpeng Xing, and Meng Han. 2026a. From retinal evidence to safe decisions: Retina-safe and ecrt for hallucination risk triage in medical llms. *arXiv preprint arXiv:2604.05348*.

Zhe Yu, Wenpeng Xing, and Meng Han. 2026b. Latentaudit: Real-time white-box faithfulness monitoring for retrieval-augmented generation with verifiable deployment. *arXiv preprint arXiv:2604.05358*.

Tianhang Zhang, Li Lin, Yelin Song, Hui Meng, Huiqiang Yin, Haifeng Li, et al. 2023a. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023b. Enhancing uncertainty-based hallucination detection with stronger focus. *arXiv preprint arXiv:2311.13230*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46.

A Metric: Pearson Correlation Coefficient (PCC) with Hallucination Severity

Beyond standard binary classification metrics (e.g., AUROC), we employ the Pearson Correlation Coefficient (PCC) to evaluate the model's capability in perceiving the *severity* of hallucinations. This metric assesses the linear alignment between the model's predicted confidence and the actual density of factual errors within a generated response.

Formally, let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the test set containing N samples. For each sample i, we define two key variables:

- **Predicted Probability (P_i):** The hallucination confidence score output by our proposed model (e.g., DISF), representing the estimated likelihood of the generated text containing hallucinations.
- **Ground Truth Severity (S_i):** The actual proportion of hallucinated content, defined as the ratio of hallucinated tokens to the total sequence length. This is calculated as:

$$S_i = \frac{L_{\text{hal}}^{(i)}}{L_{\text{total}}^{(i)}} \tag{7}$$

where $L_{\text{hal}}^{(i)}$ denotes the number of tokens annotated as hallucinations, and $L_{\text{total}}^{(i)}$ is the total number of tokens in the response.

The PCC score is then computed to measure the correlation between the prediction set $\mathbf{P} = \{P_1, \ldots, P_N\}$ and the severity set $\mathbf{S} = \{S_1, \ldots, S_N\}$:

$$\text{PCC} = \frac{\sum_i (P_i - \bar{P})(S_i - \bar{S})}{\sqrt{\sum_i (P_i - \bar{P})^2} \cdot \sqrt{\sum_i (S_i - \bar{S})^2}} \tag{8}$$

where $\bar{P}$ and $\bar{S}$ represent the mean values of the predicted probabilities and ground truth severity scores, respectively. A higher PCC indicates that the model is **well-calibrated** to the granularity of errors—assigning proportionally higher risk scores to responses that are heavily contaminated with hallucinations, rather than treating all errors equally.

B Dataset Details

B.1 RagTruth

RagTruth (Niu et al., 2024) is a large-scale RAG hallucination benchmark comprising approximately 12,000 response samples across multiple backbone LLMs.

Data Statistics.

- **Total samples:** $\sim$12,000 responses
- **Per-model distribution:** LLaMA2-7B/13B and Mistral-7B each contribute $\sim$2,965 samples; LLaMA3-8B contributes $\sim$2,950 samples
- **Train/Test split:** $\sim$2,515 training / $\sim$450 test samples per model

Task Types. The benchmark covers question answering (QA) and abstractive summarization tasks.
Input Configuration. We employ dual-path inference: (1) *CTX path*: full context with retrieved documents; (2) *NOCTX path*: query-only without retrieval context.

B.2 HalluRAG

HalluRAG (Ridder and Schilling, 2024) specifically targets knowledge-conflict scenarios where retrieved evidence contradicts the model's parametric memory.

Data Statistics.

- **Total samples:** 1,080 responses per model
- **Structure:** 540 prompts $\times$ 2 answers per prompt
- **Split ratio:** 70% train (756) / 15% validation (162) / 15% test (162)

Splitting Principle. We enforce strict *group-level splitting* to ensure that both answers from the same prompt remain in the same split, preventing content leakage across train/val/test boundaries.
Subset Selection. We use the *answerable* subset to evaluate the model's ability to prioritize retrieved context over internal priors within a closed-domain setting.

C Feature Component Details

DISF decomposes internal model dynamics into three feature families (*Conflict*, *Drift*, *Instability*), each implemented via specific computational modules:

- **Conflict** $\rightarrow$ Baseline + ICR (evidence vs. memory tension)
- **Drift** $\rightarrow$ LSD + HARP (trajectory divergence)
- **Instability** $\rightarrow$ LogitLens (internal decoherence)

This section provides implementation details for these five core components.

C.1 Baseline Features

The *Baseline* module captures fundamental token-level signals comparing CTX and NOCTX inference paths. It comprises 8 features:

- **alpha_doc / alpha_param**: Attention allocation ratios to document tokens vs. parametric tokens, measuring context reliance.
- **delta_t / delta_t_smooth**: Raw and smoothed temporal difference of hidden state norms between consecutive tokens.
- **residual_shift**: L2 distance between CTX and NOCTX hidden states at the final layer.
- **logit_shift**: $\text{Logit}(y_t|\text{CTX}) - \text{Logit}(y_t|\text{NOCTX})$, measuring how context affects token likelihood.
- **hidden_norm_ctx / hidden_norm_noctx**: L2 norms of hidden states under each path.

C.2 LSD: Layer-wise Semantic Dynamics

LSD treats transformer layers as a temporal axis and analyzes the "motion" of hidden states through representation space. The core insight is that hallucinated content exhibits erratic trajectories across layers.
Velocity Computation. Semantic velocity between consecutive layers is defined as:

$$v_\ell(t) = \|h_{t,\ell+1} - h_{t,\ell}\|_2 \tag{9}$$

where $h_{t,\ell}$ is the hidden state of token t at layer ℓ.
Acceleration Computation. Acceleration measures direction changes via cosine similarity between consecutive displacement vectors:

$$a_\ell(t) = 1 - \cos\big(\Delta h_{t,\ell}, \Delta h_{t,\ell+1}\big) \tag{10}$$

where $\Delta h_{t,\ell} = h_{t,\ell+1} - h_{t,\ell}$.
Contrastive Features. LSD computes trajectory divergence between CTX and NOCTX paths:

$$D_{\text{traj}}(t) = \|\mathbf{v}^{\text{ctx}}(t) - \mathbf{v}^{\text{noctx}}(t)\|_2 \tag{11}$$

where $\mathbf{v}(t)$ is the velocity profile across all layers.

Feature List (9 features). `max_velocity`, `late_velocity`, `total_trajectory_length`, `velocity_variance`, `max_acceleration`, `late_acceleration`, `max_velocity_layer`, `trajectory_divergence`, `velocity_correlation`.

C.3 HARP: Reasoning Subspace Projection

HARP decomposes the hidden state space into semantic and reasoning subspaces via SVD of the Unembedding matrix W_U.
Subspace Decomposition. Given $W_U = U\Sigma V^\top$, we partition the right singular vectors:

- **Semantic subspace**: Top-k singular vectors capturing 95% variance (syntax, fluency).
- **Reasoning subspace**: Remaining vectors encoding factual/logical content.

Projection Matrices.

$$P_{\text{semantic}} = V_k V_k^\top, \quad P_{\text{reasoning}} = V_{\neg k} V_{\neg k}^\top \tag{12}$$

Extracted Features (6 features).

- **reasoning_norm**: $\|P_{\text{reasoning}} h\|_2$, magnitude of reasoning component.
- **projection_ratio**: $\frac{\|P_{\text{reasoning}} h\|}{\|h\|}$, fraction in reasoning subspace.

- **semantic_norm**: $\|P_{\text{semantic}}h\|_2$.
- **reasoning_cosine_shift**: Direction change in reasoning space between tokens.
- **reasoning_delta_norm / reasoning_cosine_ctx_noctx**: Contrastive features comparing CTX vs. NOCTX reasoning projections.

C.4 ICR: Information Contribution to Residual

ICR analyzes competition between Attention (copying from context) and FFN (injecting parametric knowledge) in residual stream updates.

ICR Score. For each layer, ICR measures the divergence between:

- P_{proj}: Where the hidden state update is pointing (cosine similarity with context tokens).
- P_{attn}: Where attention is focusing (attention weight distribution).

$$\text{ICR}(\ell, t) = \text{JSD}(P_{\text{proj}}, P_{\text{attn}}) \tag{13}$$

FFN Contribution Ratio.

$$r_{\text{FFN}} = \frac{\|\Delta h_{\text{FFN}}\|}{\|\Delta h_{\text{Attn}}\| + \|\Delta h_{\text{FFN}}\|} \tag{14}$$

where $\Delta h_{\text{Attn}} = h_{\text{after-Attn}} - h_{\text{before}}$ and $\Delta h_{\text{FFN}} = h_{\text{after-FFN}} - h_{\text{after-Attn}}$.

Feature List (5 features). `icr_mean`, `icr_max`, `icr_variance`, `context_alignment_mean`, `logit_ctx_noctx_jsd`.

C.5 LogitLens: Middle Layer Decoding

LogitLens applies the final Unembedding layer to intermediate hidden states, revealing what the model "thinks" at each layer before final output.

Layer Decoding. For hidden state $h_{t,\ell}$ at layer ℓ:

$$P_\ell(y|t) = \text{softmax}(W_U \cdot \text{LayerNorm}(h_{t,\ell})) \tag{15}$$

KL Divergence Features. We compute KL divergence between intermediate and final layer predictions:

$$D_{\text{KL}}(\ell, t) = \text{KL}\left(P_\ell(y|t) \,\|\, P_L(y|t)\right) \tag{16}$$

at early ($\ell \approx 8$), mid ($\ell \approx 16$), and late ($\ell \approx 24$) layers.

Convergence Analysis.

- **first_top1_layer**: First layer where final token appears as top-1 prediction.
- **prediction_stability**: Fraction of layer pairs with consistent top-1 prediction.

Feature List (8 features). `kl_div_early`, `kl_div_mid`, `kl_div_late`, `entropy_early`, `entropy_mid`, `entropy_final`, `first_top1_layer`, `prediction_stability`.

C.6 Feature Aggregation

Each of the 36 raw features is aggregated from token-level to response-level using four pooling operators:

$$\mathbf{v}_d = \big[\text{Mean}, \text{Std}, \text{Max}, \text{TopK}\big](\mathbf{x}_d) \tag{17}$$

where TopK-Mean selects the top 10% tokens (minimum 5) by feature magnitude. This yields $36 \times 4 + 1 = 145$ features (including response length).

D Response-Level Evaluation Protocol

This section describes the DISF response-level evaluation protocol, which serves as the primary evaluation framework for DISF. The protocol is designed to provide a unified, fair, and calibrated response-level comparison.

D.1 Task Definition

Input. A model response consisting of a sequence of tokens, where each token is associated with DISF-extracted features.

Output. A scalar hallucination score per response and a binary hallucination label for evaluation.

D.2 Ground Truth Definition

Let a response consist of T tokens, each annotated with a binary hallucination label $y_t \in \{0, 1\}$.

Response-Level Label.

$$y_{\text{resp}} = \begin{cases} 1, & \text{if } \sum_t y_t > 0 \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

A response is considered hallucinated if *any* part of it contains hallucination.

Severity (for Correlation Analysis).

$$\text{severity} = \frac{1}{T}\sum_{t=1}^{T} y_t \tag{19}$$

D.3 Feature Pooling

Each token-level feature f_t is aggregated into response-level statistics using four pooling operators:

1. **Mean**: $\mu_f = \frac{1}{T}\sum_{t=1}^{T} f_t$
2. **Std**: $\sigma_f = \sqrt{\frac{1}{T}\sum_t (f_t - \mu_f)^2}$
3. **Max**: $\max_f = \max_t f_t$
4. **TopK**: Mean of top-k values, where $k = \max(k_{\min}, \lceil \rho T \rceil)$

Additionally, the response length T is appended. For F raw token-level features, the final response vector has $4F + 1$ dimensions.

D.4 Classifier and Training

Model. XGBoost binary classifier with `binary:logistic` objective.

Class Imbalance. Handled via `scale_pos_weight`.

Cross-Validation. GroupKFold by `source_id` to prevent leakage between responses from the same source.

Threshold Selection. The classification threshold τ is selected on training data using out-of-fold (OOF) predictions:

$$\tau^* = \arg\max_{\tau} \mathrm{F1}(\tau) \tag{20}$$

The same threshold is then applied to the test set.

D.5 Evaluation Metrics

Primary Metrics. AUROC, AUPRC, F1/Precision/Recall (at threshold τ^*).

Ranking Metrics. Recall@K for $K \in \{10, 50, 100\}$.

Severity Correlation. Pearson Correlation Coefficient (PCC) between predicted response score and ground-truth hallucination severity.

E Details of Baseline Methods

This section provides detailed descriptions of the eight baseline methods compared in our experiments.

E.1 Uncertainty-based Methods

These methods estimate hallucination risk via token-level probabilities or entropy.

Perplexity (Ren et al., 2023). A foundational uncertainty baseline that quantifies generation confidence via the exponentiated average negative log-likelihood of the response tokens. Higher perplexity indicates lower model confidence and potentially higher hallucination risk.

LN-Entropy (Malinin and Gales, 2020). Extends perplexity by computing length-normalized predictive entropy over the output distribution. This captures sequence-level uncertainty by aggregating token-level entropy scores, mitigating length bias in uncertainty estimation.

Focus (Zhang et al., 2023a). Enhances uncertainty estimation by focusing on information-rich tokens. It uses entropy and token probability as base scores, then calibrates them by propagating importance weights through attention distributions.

E.2 Sampling-based Methods

These methods measure consistency across multiple stochastic responses to the same query.

SelfCheckGPT (Manakul et al., 2023). A zero-resource black-box approach that detects inconsistencies via NLI over stochastic response samples. By generating multiple responses to the same query, it identifies hallucinations as claims that fail consistency checks across samples.

EigenScore (INSIDE) (Chen et al., 2024). Measures semantic inconsistency via the eigenvalues of the hidden state covariance matrix across multiple sampled responses. Large eigenvalue dispersion indicates semantic instability and potential hallucination.

E.3 Verbalization Methods

These methods prompt LLMs to self-evaluate response truthfulness or verify claims via external models.

P(True) (Kadavath et al., 2022). Prompts the model to self-evaluate its response truthfulness by asking "Is the following answer true?" The probability assigned to the "True" token serves as a confidence score. This verbalization approach leverages the model's metacognitive capabilities.

RefChecker (Hu et al., 2024). A fine-grained detector that decomposes responses into claim triplets (subject-relation-object) for entailment verification against the context. Each triplet is independently verified via an external NLI model, enabling localized hallucination detection at the claim level.

E.4 Interpretability-based Methods

These methods leverage mechanistic interpretability to decouple external context signals from internal parametric knowledge.

ReDeEP (Sun et al., 2024). A mechanistic interpretability method for RAG that detects hallucinations by monitoring when models over-rely on parametric knowledge over retrieved evidence. It decouples external context signals from internal memory signals at specific attention heads and FFN layers.

F Supervised Baseline Details

This section describes the four supervised internal-state baselines added for comparison and reports full per-backbone results under the same unified response-level protocol.

F.1 Method Descriptions

FactoScope (He et al., 2024). FactoScope detects hallucinations by decoding hidden-state evolution through Logit Lens projections across transformer layers. It constructs factual discernment features from the probability shift patterns of intermediate-layer predictions, capturing when the model's internal factual representations diverge from its final output. Under our protocol, we extract FactoScope-style features for each backbone and train the same XGBoost readout.

UQ Heads (Kossen et al., 2024). UQ Heads trains auxiliary uncertainty quantification probes on attention head activations. It identifies attention patterns that correlate with hallucination risk, using a subset of attention heads as uncertainty estimators. We adapt UQ-Head features to our response-level evaluation pipeline.

LookbackLens (Chuang et al., 2024). LookbackLens measures attention weights allocated to context tokens versus parametric tokens across layers. Originally designed for context-grounding analysis, we adapt it to the RAG hallucination detection setting by computing context-attention ratio features at each token position and applying the same aggregation and readout pipeline.

SpikeScore (Liang et al., 2024). SpikeScore exploits cross-layer entropy spikes in hidden-state predictions for hallucination detection. It identifies layers where the predictive entropy exhibits sudden spikes, treating these as indicators of internal uncertainty or knowledge conflict. Notably, SpikeScore was designed for cross-domain detection, making it a particularly relevant baseline for evaluating DISF's generalization claims.

F.2 Per-Backbone Results

Table 4: Full per-backbone comparison of supervised internal-state baselines. All methods use the same response-level protocol (GroupKFold split, XGBoost readout). **Bold** = best per setting.

		RagTruth		HalluRAG	
Backbone	**Method**	AUROC↑	AUPRC↑	AUROC↑	AUPRC↑
LLaMA2-7B	SpikeScore	0.7308	0.7727	0.7876	0.7658
	FactoScope	0.7812	0.7989	0.9480	0.9460
	UQ Heads	0.7580	0.7950	0.8120	0.8030
	LookbackLens	0.6820	0.6950	0.7250	0.7100
	DISF	**0.7947**	**0.8133**	**0.9582**	**0.9573**
LLaMA2-13B	SpikeScore	0.7381	0.7677	0.7912	0.8014
	FactoScope	0.8450	0.8410	0.8540	0.8920
	UQ Heads	0.7820	0.8050	0.8130	0.8340
	LookbackLens	0.7050	0.7230	0.7180	0.7310
	DISF	**0.8578**	**0.8558**	**0.8655**	**0.9027**
LLaMA3-8B	SpikeScore	0.7809	0.7863	0.8261	0.7972
	FactoScope	0.9250	0.8750	0.8380	0.8350
	UQ Heads	0.8320	0.8280	0.8040	0.7950
	LookbackLens	0.7280	0.7420	0.6940	0.7320
	DISF	**0.9354**	**0.8849**	**0.8489**	**0.8410**
Mistral-7B	SpikeScore	0.7676	0.7322	0.7981	0.7542
	FactoScope	0.8340	0.8490	0.9170	0.9370
	UQ Heads	0.7830	0.8260	0.7940	0.7990
	LookbackLens	0.6920	0.7050	0.7100	0.7770
	DISF	**0.8448**	**0.8641**	**0.9246**	**0.9416**

F.3 XGBoost vs. Linear Readout

To isolate the effect of the downstream classifier, we compare XGBoost with a simpler logistic regression (LR) readout for both DISF and supervised baselines (Table 5).

Table 5: XGBoost vs. logistic regression readout (8-setting average). DISF's advantage persists with the simpler readout.

Method	**Readout**	AUROC↑	AUPRC↑	PRR↑
DISF (Full)	XGBoost	**0.8891**	**0.8986**	**0.7830**
FactoScope	XGBoost	0.8779	0.8826	0.7758
UQ Heads	XGBoost	0.7973	0.8231	0.5907
LookbackLens	XGBoost	0.7068	0.7269	0.4787
DISF (Full)	LR	**0.8801**	**0.8720**	**0.7757**
FactoScope	LR	0.8762	0.8708	0.7678
UQ Heads	LR	0.7950	0.8128	0.5919
LookbackLens	LR	0.7191	0.7255	0.4778

G Supplementary Experimental Results

This section consolidates additional experimental evidence referenced in the main text, including generalization to newer backbones, cross-model transfer, decoding robustness, deployment-oriented metrics, and multi-metric stability analysis.

G.1 Qwen3 Backbone Results

To validate DISF on newer-generation models, we evaluate on `Qwen3-8B` and `Qwen3-14B` under the same response-level protocol (Table 6).

Table 6: DISF results on Qwen3 backbones (same protocol as main experiments).

Backbone	Dataset	AUROC↑	AUPRC↑	PCC↑
Qwen3-8B	RagTruth	0.8615	0.8526	0.3796
	HalluRAG	0.8768	0.8840	0.5612
Qwen3-14B	RagTruth	0.8443	0.7169	0.4432
	HalluRAG	0.9146	0.7659	0.5944

Results confirm that DISF generalizes well to newer instruction-tuned architectures. `Qwen3-14B` achieves particularly strong detection on *HalluRAG* (AUROC 0.9146), suggesting that dual-path contrastive signals remain informative even for models with stronger internal alignment.

G.2 LOMO Cross-Model Transfer

We evaluate leave-one-model-out (LOMO) cross-model transfer: train DISF on three backbones, test on the held-out backbone under the same response-level protocol. This tests whether DISF features generalize across model families without target-model training.

Table 7: LOMO cross-model transfer results (AUROC / AUPRC). Train on 3 backbones, test on the held-out backbone.

	RagTruth		HalluRAG	
Held-out Backbone	AUROC	AUPRC	AUROC	AUPRC
LLaMA2-7B	0.7848	0.7743	0.8770	0.8207
LLaMA2-13B	0.7968	0.7527	0.7735	0.7604
LLaMA3-8B	0.8477	0.7830	0.7600	0.8888
Mistral-7B	0.7519	0.7602	0.7755	0.6765
Mean	0.7953	0.7675	0.7965	0.6866
Overall Mean (8 settings)	AUROC: 0.7959		AUPRC: 0.7271	

The LOMO results demonstrate that DISF's internal-state features capture cross-model hallucination patterns without requiring target-backbone supervision. This supports a practical *cost-saving deployment mode*: when fast rollout is needed, a pre-trained multi-model detector can be deployed directly on a new backbone with meaningful performance.

G.3 Decoding Sensitivity Analysis

Following reviewer suggestions, we test whether DISF signals collapse under stronger decoding regimes. We compare two settings on *HalluRAG*: (1) t=0.7, p=0.9 (moderate) vs. (2) t=1.0, p=0.95 (permissive), with separately regenerated and relabeled data for each setting.

Table 8: Decoding sensitivity analysis on HalluRAG. Values should be compared within-analysis (t07p09 vs. t10p095), not against main-table absolutes. No signal collapse is observed.

	t07p09			t10p095			Δ (t10−t07)		
Backbone	AUC	PRC	PCC	AUC	PRC	PCC	AUC	PRC	PCC
Qwen3-14B	0.9332	0.7779	0.6957	0.9147	0.7267	0.6394	−0.019	−0.051	−0.056
Qwen3-8B	0.8574	0.8623	0.5291	0.8219	0.8136	0.4813	−0.036	−0.049	−0.048
LLaMA2-7B	0.9386	0.9462	0.6552	0.8914	0.8963	0.5671	−0.047	−0.050	−0.088
LLaMA2-13B	0.8386	0.9079	0.5103	0.7933	0.8597	0.4481	−0.045	−0.048	−0.062
LLaMA3-8B	0.8455	0.8441	0.5482	0.8212	0.7848	0.4947	−0.024	−0.059	−0.054
Mistral-7B	0.8368	0.8476	0.3762	0.7896	0.8067	0.2871	−0.047	−0.041	−0.089

Key findings: (1) All models remain clearly above random ranking under both decoding settings (AUROC 0.7896–0.9332). (2) The moderate degradation under permissive decoding is model-dependent, with no universal signal collapse. (3) Qwen3-14B shows the least sensitivity, maintaining 0.9147 AUROC even under t10p095.

G.4 Selective Prediction Metrics

Beyond ranking-based metrics, we report deployment-oriented selective prediction metrics: Prediction-Rejection Ratio (PRR), Area Under the Risk-Coverage curve (AURC), and Risk at 80% Coverage (Risk@80%Cov). Averaged over 8 settings:

Table 9: Selective prediction metrics (8-setting average, XGBoost readout).

Metric	DISF	FactoScope	UQ Heads
PRR↑	**0.7830**	0.7758	0.5907
AURC↓	**0.2634**	0.2720	0.3150
Risk@80%↓	**0.4308**	0.4380	0.4650

DISF wins Risk@80%Cov on 7 out of 8 settings, confirming its superiority in practical deployment scenarios where a fixed coverage target must be met.

G.5 Multi-Metric Stability Analysis

To address concerns about single-metric cherry-picking, we compute the average rank of each ablation variant across all metrics and all 8 settings, providing a holistic view of feature importance.

Table 10: Multi-metric average-rank stability analysis (lower = more stable). Computed across AUROC, AUPRC, PRR, AURC, Risk@80%Cov × 8 settings.

Ablation Variant	Avg Rank↓
Full DISF	**3.562**
w/o Conflict	4.125
w/o Drift	6.391
w/o Instability	5.438
Only Conflict	8.703
Only Drift	5.812
Only Instability	9.438

Key insight: Full DISF achieves the most stable performance envelope (avg_rank 3.562). Removing Drift severely degrades stability (avg_rank 6.391) and selectively damages vital deployment metrics (AURC worsening from 0.2634 to 0.2824). This confirms that all three feature families contribute complementary detection signals, and Drift serves as a particularly critical universal component.

H Practical Deployment Recipes

To concretely address the latency overhead of dual-path inference in real-world RAG systems, we propose three flexible deployment strategies:

Cascade Mode. RAG systems can run a lightweight single-path filter first (e.g., the *Instability-only* configuration, which adds only 43.7% overhead). Full dual-path DISF is invoked only for responses flagged as uncertain or high-risk, amortizing the cost across the traffic mix.

Cache-Aware Reuse. Since the CTX path exactly mirrors the primary generation pass, efficient implementations can cache the CTX hidden states and logits directly during generation. This limits the marginal cost of DISF entirely to the deterministic NOCTX teacher-forced replay, effectively halving the structural overhead.

Batched NOCTX Replay. For high-volume deployment, NOCTX replays across multiple independent responses can be batched together on the GPU, significantly boosting global throughput compared to sequential per-response replay.

These strategies enable DISF to serve interchangeably as either a lightweight real-time guardrail or a robust secondary auditor for high-stakes reasoning, depending on the quality–latency requirements of the deployment scenario.

Table 11: Comprehensive ablation study of DISF across backbone models and datasets (*RagTruth* and *HalluRAG* test sets). "w/o X" denotes removal of feature group X; "Only X" denotes using only that group. **Bold** and underline indicate the best and second-best results *within each model*, respectively.

			RagTruth (Test Set)				HalluRAG (Test Set)			
Model	**Variant**	**#Feat**	**AUROC↑**	**AUPRC↑**	**F1↑**	**PCC↑**	**AUROC↑**	**AUPRC↑**	**F1↑**	**PCC↑**
LLaMA2-7B	**Ours (Full)**	145	**0.7947**	**0.8133**	0.6854	0.2043	0.9582	**0.9573**	0.9167	0.6880
	w/o Conflict	93	0.7821	0.7950	0.6935	0.1899	0.9512	0.9427	0.9059	0.6865
	w/o Drift	85	0.7729	0.7993	0.6851	**0.2202**	0.8855	0.8957	0.8121	0.5573
	w/o Instability	113	0.7878	0.8066	0.6948	0.1488	**0.9584**	0.9555	0.9112	**0.6966**
	w/o ICR	125	0.7807	0.8021	**0.7071**	0.1918	0.9555	0.9547	**0.9176**	0.6842
	w/o Baseline	113	0.7817	0.7948	0.7012	0.2001	0.9497	0.9447	0.9112	0.6813
	w/o HARP	121	0.7744	0.7958	0.7014	0.1963	0.9413	0.9222	0.8970	0.6550
	w/o LSD	109	0.7699	0.8047	0.6779	0.1843	0.9186	0.9268	0.8280	0.6043
	w/o LogitLens	113	0.7734	0.8066	0.6948	0.1488	**0.9584**	0.9555	0.9112	**0.6966**
	Only Conflict	53	0.7210	0.7228	0.6686	0.0646	0.8828	0.8934	0.7805	0.5516
	Only Drift	61	0.7583	0.7715	0.6768	0.1321	0.9515	0.9478	0.8889	0.6952
	Only Instability	33	0.7186	0.7359	0.6687	0.1765	0.8078	0.7981	0.7684	0.4738
LLaMA3-8B	**Ours (Full)**	145	**0.9354**	0.8849	0.8475	0.3849	**0.8489**	0.8410	0.7850	0.5840
	w/o Conflict	93	0.9341	0.8612	0.8475	0.3887	0.8035	0.7920	0.7240	0.4820
	w/o Drift	85	0.9047	0.8307	0.7937	0.3712	0.8088	0.8150	**0.7960**	0.6210
	w/o Instability	113	0.9061	0.8344	0.7797	0.3491	0.8134	0.7980	0.7150	0.5230
	w/o ICR	125	0.9318	0.8901	0.7931	0.3790	0.8060	0.7850	0.7620	0.5140
	w/o Baseline	113	0.9332	**0.8931**	0.8276	0.3549	0.8032	0.7940	0.6850	0.4620
	w/o HARP	121	0.9250	0.8709	0.8475	0.3637	0.8278	0.8120	0.7540	0.5360
	w/o LSD	109	0.9079	0.8729	0.8000	0.3729	0.7701	0.7560	0.6620	0.4850
	w/o LogitLens	113	0.9061	0.8344	0.7797	0.3491	0.8134	0.7980	0.7150	0.5230
	Only Conflict	53	0.9313	0.8908	**0.8621**	**0.3889**	0.8160	**0.8520**	**0.8120**	**0.6850**
	Only Drift	61	0.9341	0.8828	0.8065	0.3728	0.7616	0.7420	0.6580	0.4120
	Only Instability	33	0.8943	0.8839	0.7458	0.3220	0.7171	0.6850	0.6120	0.3950
LLaMA2-13B	**Ours (Full)**	145	**0.8578**	0.8558	**0.7780**	0.2281	0.8655	0.9027	0.7816	0.5067
	w/o Conflict	93	0.8507	0.8464	0.7644	**0.2284**	0.8641	0.8975	**0.8068**	0.4921
	w/o Drift	85	0.8539	**0.8604**	0.7529	0.1931	0.8395	0.8905	0.7654	0.4566
	w/o Instability	113	0.8177	0.7951	0.7512	0.1763	0.8770	0.9076	0.8023	**0.5317**
	w/o ICR	125	0.8518	0.8483	0.7766	0.2235	0.8621	0.9009	0.7791	0.4990
	w/o Baseline	113	0.8423	0.8415	0.7696	0.2223	0.8738	0.8982	0.7879	0.5191
	w/o HARP	121	0.8459	0.8438	0.7677	0.2088	0.8753	0.9082	0.7976	0.5152
	w/o LSD	109	0.8552	0.8568	0.7345	0.2064	0.8365	0.8843	0.7602	0.4679
	w/o LogitLens	113	0.8177	0.7951	0.7512	0.1763	0.8770	0.9076	0.8023	**0.5317**
	Only Conflict	53	0.8217	0.8212	0.7309	0.1787	0.8409	0.8731	0.7500	0.4564
	Only Drift	61	0.8421	0.8148	0.7472	0.2021	**0.8957**	**0.9109**	0.7976	0.5257
	Only Instability	33	0.8317	0.8440	0.7261	0.1664	0.7399	0.8115	0.7059	0.3420
Mistral-7B	**Ours (Full)**	145	0.8448	**0.8641**	0.7879	0.2162	0.9246	0.9416	0.8605	0.6681
	w/o Conflict	93	0.8391	0.8476	0.7797	0.2042	0.9295	0.9418	0.8538	**0.6853**
	w/o Drift	85	0.8391	0.8488	0.7783	0.1958	0.9004	0.9168	0.8166	0.6229
	w/o Instability	113	**0.8458**	0.8538	**0.8132**	0.2072	0.9158	0.9338	0.8523	0.6417
	w/o ICR	125	0.8396	0.8474	0.7877	0.2192	**0.9300**	**0.9452**	**0.8639**	0.6800
	w/o Baseline	113	0.8353	0.8441	0.7880	0.2023	0.9235	0.9384	0.8452	0.6744
	w/o HARP	121	0.8333	0.8422	0.7787	0.2025	0.9229	0.9375	0.8471	0.6535
	w/o LSD	109	0.8265	0.8176	0.7814	0.1944	0.9230	0.9388	0.8555	0.6607
	w/o LogitLens	113	**0.8458**	0.8538	**0.8132**	0.2072	0.9158	0.9338	0.8523	0.6417
	Only Conflict	53	0.8011	0.8065	0.7512	0.1788	0.8312	0.8744	0.7261	0.5340
	Only Drift	61	0.8446	0.8472	0.7891	**0.2206**	0.9180	0.9342	0.8383	0.6639
	Only Instability	33	0.8187	0.8289	0.7754	0.1715	0.7930	0.8062	0.7778	0.5034