HETEROGENEOUS UNCERTAINTY ALONG DENOISING TRACES SIGNALS HALLUCINATIONS IN DIFFUSION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Given the rich token-confidence information they contain across positions and time steps, the denoising traces of diffusion language models offer a natural foundation for uncertainty quantification and hallucination detection. Yet, existing detection methods either derive from autoregressive models—failing to capture the full extent of this information—or process it using heavyweight neural architectures. In this paper, we show that hallucinations are better predicted by the heterogeneity of uncertainty along this trace than by its average level. In particular, large variations in token distribution entropy over denoising steps provide a strong discriminative signal. A Markovian model of token-wise convergence links this pattern to slower and more heterogeneous convergence in hallucinated answers. Building on these findings, we introduce DRIFTY: a lightweight detector combining 37 closed-form trace features with regularized logistic regression. Across three diffusion language models and three closed-book question-answering benchmarks, DRIFTY outperforms all evaluated baselines in every model–dataset pair, improving over Semantic Entropy by up to 14.5 ROC-AUC points and 16 PR-AUC points. Moreover, DRIFTY uses only the trace of a single deterministic generation, with feature extraction and classification requiring roughly half a second of CPU time per 100 samples.

1 INTRODUCTION

Diffusion language models (d-LLMs) such as LLaDA (Nie et al., 2025; Zhu et al., 2026), Dream (Ye et al., 2025) and DiffusionGemma (DiffusionGemma Team et al., 2026) have emerged as credible alternatives to autoregressive (AR) LLMs: instead of emitting tokens left-to-right, they reconstruct the answer by iteratively denoising a fully masked sequence with bidirectional attention, enabling flexible order, parallel decoding (Ben-Hamu et al., 2026), and revision of already-produced tokens. Like AR models, they hallucinate - fluent, confident, factually wrong answers - but hallucination detection has been studied almost exclusively in the AR setting (Ji et al., 2023; Liu et al., 2025).

Established AR detectors assume uncertainty is a one-dimensional object: one distribution per token, produced once, in a fixed order. Perplexity (Jelinek et al., 1977) and length-normalized entropy (Malinin & Gales, 2021) average token log-probabilities and entropies; semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024) clusters multiple samples by entailment; lexical similarity (Fomicheva et al., 2020) measures ROUGE-L consistency across re-samples. Transposed to d-LLMs, these methods can only read each token *after the fact*, at the step it is fixed, and discard the rest of the denoising history.

Yet a d-LLM produces a richer object at inference: at every step t and position d it re-evaluates its full predictive distribution, yielding a $T \times D$ set, called *trace*, of log-probabilities, entropy values and masks. This trace carries hallucination signatures that the final answer alone does not. As an illustration, Chang et al. (2025) distinguish interleaved spurious content, oscillation between incompatible guesses, and early lock-in on a wrong answer (persistent error); what these profiles share is *hesitation* — not the average level of uncertainty, but how unevenly it is distributed over positions and time.

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.

	LLaDA-8B	Dream-7B	DiffusionGemma
Parameters	8B (trained from scratch)	7B (init. Qwen 2.5 / LLaMA 3)	26B MoE (3.8B active)
Diffusion	Masked, absorbing	Masked, absorbing	Uniform, re-noiseable
Remasking	Low-confidence	Low-confidence	Low-confidence variant (native: entropy-bounded)
Canvas	Fixed length	Fixed length	Resized to gen. length (native 256)

Table 1: The three diffusion language models as used in this paper. DiffusionGemma entries describe our protocol; native defaults in parentheses (see Section 4).

Existing approaches exploit this trace using complex neural architectures (Chang et al., 2025; Hemmat et al., 2026) or zero-shot approximations (Vazhentsev et al., 2026). However, these neural detectors require dedicated multi-parameter training and substantial inference overhead (up to hundreds of seconds per 100 samples). This complexity raises a fundamental question: does hallucination detection truly require heavy learned representations of the trajectory, or can the underlying phenomenon be captured much more simply? We formulate this as our central question: **what is a sufficient, low-dimensional summary of the denoising trace for hallucination detection?**

We answer with DRIFTY, a deliberately lightweight pipeline: 37 closed-form trace statistics (confidence level, cross-token dispersion and shape of the heterogeneity curve, Markovian convergence rates) fed to an L_1/L_2-regularized logistic regression as the only learned component. The design follows one structural insight, grounded in a per-token Markovian convergence model: since all tokens start near-uniform and end near-deterministic, mean-level dynamics mostly reflect the diffusion schedule, while cross-token heterogeneity directly encodes dispersed convergence speeds and separates hallucinations from correct answers. A complementary Ornstein–Uhlenbeck (OU) analysis yields an instructive negative result, ruling out temporal persistence as the signal.

Evaluated on three d-LLMs (LLaDA-8B, Dream-7B, DiffusionGemma-26B) and three closed-book QA benchmarks (TriviaQA, HotpotQA, Natural Questions), DRIFTY dominates all four AR baselines in every model–dataset pair using a single deterministic generation and CPU-only post-processing, matches or exceeds the diffusion-specific TraceDet in most cells at a fraction of its cost, generalizes across datasets within a few points, and shows a mechanism-consistent drop on uniform-state diffusion.

Our contributions are: **(1) Lightweight detector:** closed-form statistics plus regularized logistic regression; no neural detector, one deterministic pass, and sub-second CPU scoring (Section 3–4). **(2) Inhomogeneity over level:** dispersion features dominate label correlation and ablations, backed by a Markovian rate model explaining slower hallucination convergence, while mean-level OU/AR(1) dynamics prove non-discriminative (Section 3). **(3) Comprehensive evaluation:** consistent gains across 3 d-LLMs and 3 QA benchmarks, systematic improvements over AR and diffusion baselines, cross-dataset transfer, and mechanism-consistent analysis on uniform diffusion (Section 5).

2 BACKGROUND AND RELATED WORK

2.1 DIFFUSION LANGUAGE MODELS

Discrete diffusion models (Sohl-Dickstein et al., 2015; Hoogeboom et al., 2021) corrupt data over discrete states via Markovian transition matrices; D3PM (Austin et al., 2021) generalized this framework with structured categorical corruptions. Among these schemes, the *absorbing* one - a corrupted token is replaced by a special `[MASK]` state - proved best suited to language and established the link with masked language modeling. Masked diffusion language models (MDLM, Sahoo et al., 2024) specialize this scheme and simplify the training objective to a re-weighted sum of cross-entropy losses on masked positions, using a bidirectional (non-causal) Transformer as the denoiser. LLaDA (Nie et al., 2025) and Dream (Ye et al., 2025) instantiate this paradigm at scale, whereas DiffusionGemma (DiffusionGemma Team et al., 2026) uses *uniform-state* diffusion with iterative re-noising. Table 1 summarizes the three models as used in this paper; their mechanistic differences matter for our analysis in Section 5.2.

2.2 HALLUCINATION DETECTION FOR AUTOREGRESSIVE LLMS, AND ITS TRANSPOSITION

We consider four widely adopted baselines from the AR literature, spanning single-pass and multi-sample paradigms. *Perplexity* (Jelinek et al., 1977) and *length-normalized entropy* (Malinin & Gales, 2021) are single-pass: they average token log-probabilities and predictive entropies along the generated sequence. *Semantic entropy* (Kuhn et al., 2023; Farquhar et al., 2024) and *lexical similarity* (Fomicheva et al., 2020) are multi-sample: the former clusters $N = 5$ samples (temperature 0.5) by bidirectional NLI entailment and reports cluster entropy; the latter aggregates ROUGE-L overlap between a candidate and $N - 1$ alternatives into consistency features for a logistic head.

Because a d-LLM generates sequences non-autoregressively, transposing these AR metrics requires adapting to an iterative, bidirectional trajectory. Under deterministic decoding (temperature $T = 0$, low-confidence remasking), the denoising schedule defines an empirical unmasking order π. Perplexity and LN-entropy can then be transposed by evaluating each token's log-probability and predictive entropy strictly at its fixation step. While standard, this transposition is inherently lossy: it records only a single scalar snapshot per position and discards the full $T \times D$ trajectory that produced it.

2.3 TRACE-BASED DETECTORS FOR D-LLMS

To overcome the lossy nature of transposed AR metrics, recent methods exploit the full spatio-temporal trajectory produced during diffusion generation. TraceDet (Chang et al., 2025) formalizes denoising as an MDP and extracts an informative sub-trace via an information-bottleneck Transformer. TDGNet (Hemmat et al., 2026) constructs a temporal dynamic graph from attention maps, combining per-token GRU memory with temporal attention to predict hallucinations. However, both approaches rely on dedicated neural detectors tuned over multiple hyperparameters, incurring substantial computational overhead (reported at 147.5 s and 442 s per 100 samples, respectively). Because TDGNet released no public code, we cite its reported figures as qualitative reference. For TraceDet, where the public repository departs in minor architectural dimensions from the paper, we conduct a faithful reproduction using the official repository defaults on our exact traces and splits (details in Section 4).

While we share their core premise that factuality signatures are distributed along the trajectory, we replace learned neural architectures with closed-form statistics operating within the cost class of single-pass AR baselines.

In parallel, Vazhentsev et al. (2026) conduct a systematic zero-shot study of uncertainty quantification for d-LLMs, combining level-type trajectory statistics (such as average predictive entropy, commitment-step log-probabilities, and remasking counts) with cross-encoder semantic dissimilarity between intermediate states and the final output. While their scores are unsupervised aggregates focused on mean confidence levels, DRIFTY demonstrates that it is specifically the *cross-token dispersion* along the trace — rather than its average level — that provides the primary discriminative signal for hallucination detection.

3 METHOD: HESITATION IN THE DENOISING TRACE

3.1 THE DENOISING TRACE AS A TWO-DIMENSIONAL UNCERTAINTY OBJECT

Consider a d-LLM generating an answer of length D in T denoising steps. At step t and position d, let $H_{t,d}$ denote the Shannon entropy of the model's predictive distribution over the vocabulary, $\ell_{t,d}$ the log-probability it assigns to the token that will eventually occupy position d, and $m_{t,d} \in \{0, 1\}$ the mask indicator (1 if position d is still masked at step t). The *denoising trace* is the collection $(H, \ell, m) \in (\mathbb{R} \times \mathbb{R} \times \{0, 1\})^{T \times D}$. While an AR LLM yields one scalar pair per token, a d-LLM natively produces the full $T \times D$ trajectory. We aim to reduce this object to compact statistics sufficient for hallucination detection.

All generations in this paper use temperature-zero decoding with low-confidence remasking, which makes the trajectory deterministic: at each step, the model predicts distributions for all masked positions, and only the most confident ones are fixed (for DiffusionGemma, a non-recommended deterministic variant with canvas length set to generation length; see Section 4). This is the standard

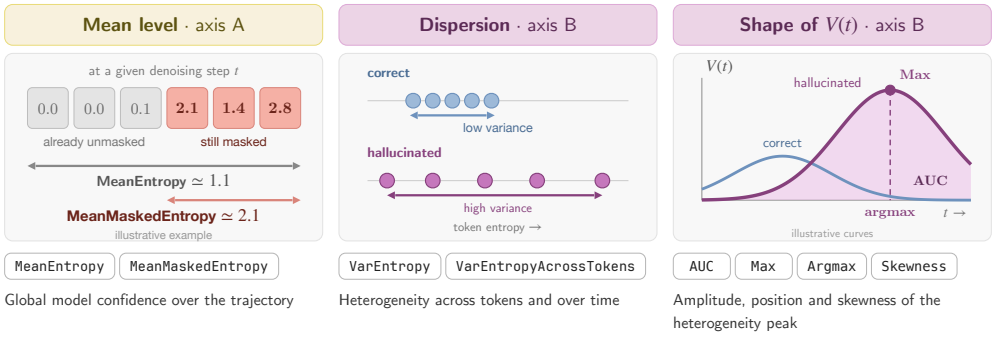

Figure 1: Overview of DRIFTY: the trace (H, ℓ, m) is distilled into 37 closed-form features (level, dispersion, dynamics, and Markovian rates), mapped by regularized logistic regression to a hallucination score. No neural detector is trained.

setting in which AR baselines can be transposed (Section 2.2), so all methods below see exactly the same trajectories.

3.2 LEVEL VERSUS INHOMOGENEITY

Two readings of the trace (H, ℓ, m) are possible. *Level* features average confidence over the trajectory (`MeanEntropy`, `MeanMaskedEntropy`, `MeanLogProbMasked` and variants). The masked restriction is essential: a fixed token has near-zero entropy, so averaging over the full $T \times D$ grid dilutes the signal with trivial zeros.

The core of DRIFTY is the second reading: *inhomogeneity*. Its central statistic is the inter-token variance among still-masked positions at step t,

$$V(t) \;=\; \mathrm{Var}_d(H_{t,d} \mid m_{t,d} = 1)\,, \tag{1}$$

with shape descriptors of $t \mapsto V(t)$ (area, max, argmax, skewness, kurtosis, mean curvature). $V(t)$ is structurally pinned near zero at both ends: all tokens start near-uniform ($H \approx \ln|\mathcal{V}|$) and end fixed at near-zero entropy. Under perfect homogeneity it would vanish identically; the mid-trajectory bell it forms is thus by construction a measure of dispersed per-token convergence speeds (see Figure 1 for a schematic view).

Figure 2 shows two TriviaQA traces (Dream, 16 steps $\times$ 32 positions): the correct answer converges almost simultaneously, while the hallucinated one keeps a wide band at high entropy for most of the run. The early low-entropy band on the right corresponds to trivial padding tokens (`<endoftext>`) that are fixed first. At population level (Figure 3, 2,500 samples with relatively balanced classes in total), the masked mean separates mid-trajectory (≈ 2.2 correct vs. ≈ 2.9 hallucinated) then rises toward the end due to a low-confidence remasking selection effect, as the shrinking pool of masked positions contracts toward the hardest tokens. $V(t)$ separates more cleanly by shape: early low peak for correct answers (step ≈ 1, amplitude ≈ 2.3) versus late high peak for hallucinations (steps 12–13, amplitude ≈ 5.2). Mean and variance are complementary views of one heterogeneity; we keep both.

3.3 A MARKOVIAN MODEL OF PER-TOKEN CONVERGENCE

To formalize the intuition behind trace heterogeneity, we model the predictive distribution at position d as converging toward its final token $i_d^* \in \{1, \ldots, |\mathcal{V}|\}$. Specifically, we assume probability mass $1 - \varepsilon_d(t)$ is allocated to i_d^*, while the remaining mass $\varepsilon_d(t)$ is distributed uniformly across the other $|\mathcal{V}| - 1$ vocabulary entries, where

$$\varepsilon_d(t) \;=\; C_d\, e^{-t/\tau_d}, \tag{2}$$

with $\tau_d > 0$ denoting the position's characteristic time constant (reflecting token difficulty) and $C_d > 0$ a boundary constant. Under this formulation, the predictive entropy $S_d(t)$ at position d

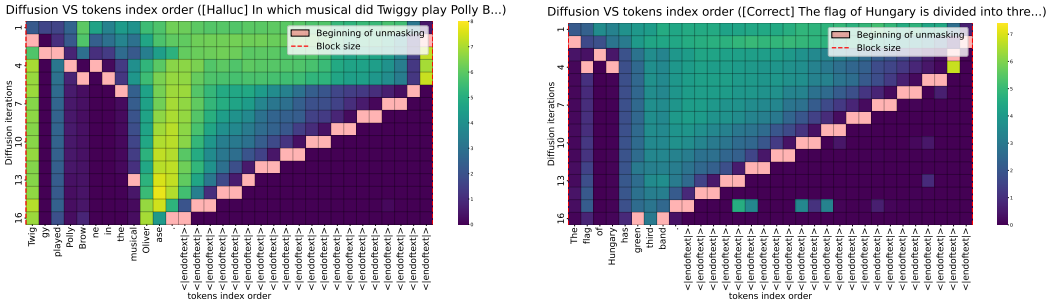

Figure 2: Token-wise entropy $H_{t,d}$ along the denoising trace (Dream, TriviaQA, 16 steps $\times$ 32 positions, low-confidence remasking). **Left:** hallucinated answer - a wide band of positions stays at high entropy for most of the trajectory. **Right:** correct answer - positions converge almost simultaneously to near-zero entropy.

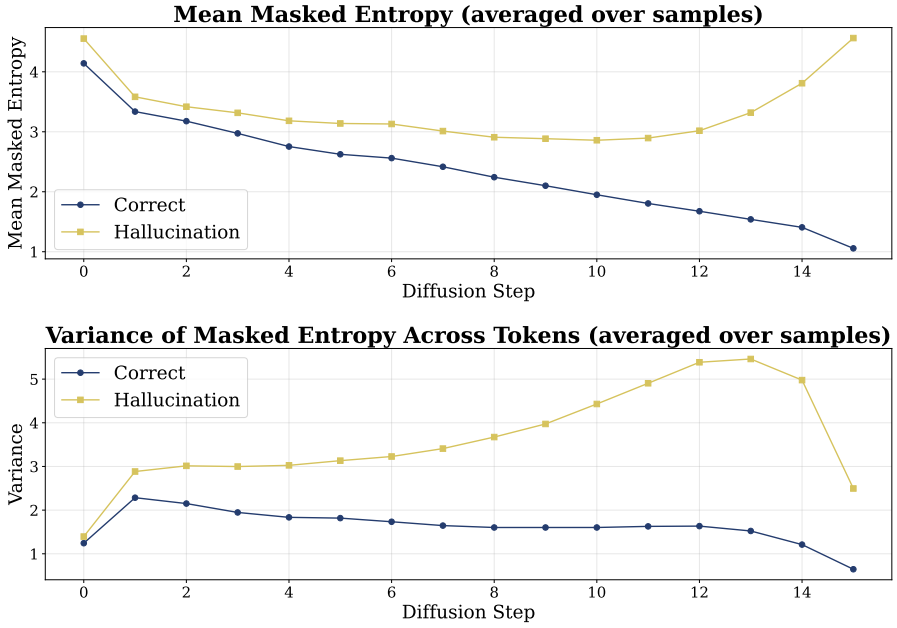

Figure 3: Masked mean entropy (top) and inter-token variance $V(t)$ (bottom), averaged over 2,500 TriviaQA samples (Dream, 16 steps, 32 tokens). Correct answers converge early and homogeneously; hallucinations converge late and heterogeneously, producing a later and higher variance peak.

follows (see Appendix A for the complete derivation):

$$S_d(t) \;\approx\; C_d\, e^{-t/\tau_d} \left(\frac{t}{\tau_d} + \ln \frac{|\mathcal{V}| - 1}{C_d} + 1 \right). \tag{3}$$

This expression yields an asymmetric profile: entropy detaches from the initial uniform prior before decaying exponentially as confidence solidifies. For a sequence of K tokens with heterogeneous rates, the average sequence entropy $\langle S \rangle(t) = \frac{1}{K} \sum_d S_d(t)$ is asymptotically governed by the slowest positions at large t, and can be parameterized as:

$$\langle S \rangle(t) \;\approx\; C\,(t - m)\, \exp\!\left(-\frac{t - m}{\tau_{\text{eff}}} \right), \tag{4}$$

where τ_{eff} represents the effective sequence-level timescale and m is a pre-peak delay. Because all tokens share near-identical initial and terminal boundary conditions, the cross-token variance $V(t)$

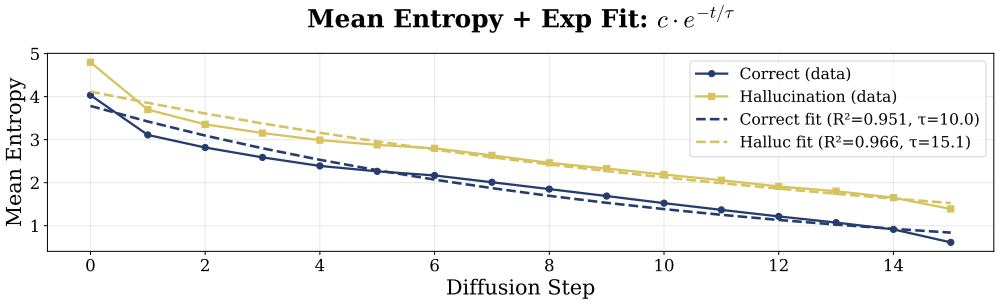

Figure 4: Mean entropy per class (2,500 TriviaQA samples, Dream, 16 steps, 32 tokens) with exponential fits $\propto e^{-t/\tau}$: $\hat{\tau} = 10.0$ correct ($R^2 = 0.951$) vs. $\hat{\tau} = 15.1$ hallucinated ($R^2 = 0.966$).

remains bounded near zero at both ends. Its intermediate bell shape therefore directly reflects the *dispersion* of individual time constants τ_d, rather than the trajectory of any single token. Consequently, hallucinations exhibit slower and more disparate dynamics, resulting in larger, more dispersed τ_d values that manifest as a delayed, higher, and more asymmetric $V(t)$ peak.

To empirically validate this behavior, we fit exponential decays $\propto e^{-t/\tau}$ to the log mean entropy $\ln \overline{\langle S \rangle}(t)$ across class-averaged Dream traces (Figure 4). The estimated time constants yield $\hat{\tau} = 10.0$ for correct answers versus $\hat{\tau} = 15.1$ for hallucinations, confirming a $\sim$50% slower convergence rate for hallucinated sequences. We fit unmasked means because low-confidence remasking introduces non-monotonic selection effects in masked entropy, violating log-linearity. These dynamics directly motivate our Markovian feature set: per-token log-linear slopes $\alpha \approx -1/\tau_d$ of $\ln H_{t,d}$, sequence-level slope descriptors $\ln\langle S \rangle(t)$, and the parametric $(C, \tau_{\text{eff}}, m)$ estimates.

3.4 Ornstein–Uhlenbeck refinement

While the deterministic envelope (3) describes the macroscopic decay of entropy, empirical trajectories exhibit local stochastic fluctuations throughout denoising. To capture these dynamics, we model token entropy as a continuous-time stochastic process $X_{t,d}$ governed by an inhomogeneous Ornstein–Uhlenbeck (OU) process (Gardiner, 2009):

$$dX_{t,d} = \mu_d(t)\big(\theta_d(t) - X_{t,d}\big)\, dt + \sigma_d(t)\, dW_{t,d}, \tag{5}$$

where $\mu_d(t) \sim 1/\tau_d$ governs the mean-reversion speed toward a time-dependent target $\theta_d(t)$, and $W_{t,d}$ denotes standard Brownian motion. At unit step increments ($\Delta t = 1$), this continuous process discretizes exactly to an autoregressive AR(1) model:

$$X_{t,d} = \varphi_{t,d}\, X_{t-1,d} + c_{t,d} + \sigma^{\varepsilon}_{t,d}\, \varepsilon_{t,d}, \qquad \varepsilon_{t,d} \sim \mathcal{N}(0,1), \tag{6}$$

with transition parameters $\varphi_{t,d} = e^{-\mu_d(t)}$ and $c_{t,d} = \theta_d(t)(1 - e^{-\mu_d(t)})$. For each class (correct vs. hallucinated) and coordinate (t, d), we estimate these coefficients by fitting ordinary least squares (OLS) of $X_{t,d}$ on $X_{t-1,d}$ pooled across all class sequences. Because each regression is strictly univariate ($p = 1 \ll n$), the resulting estimators are statistically stable and immune to overfitting.

This refinement yields an instructive negative result (see Appendix B for detailed empirical trajectories). Synthetic paths sampled from AR(1) models fitted separately on correct and hallucinated traces produce nearly superposable mean and variance trajectories. Rather than revealing a distinctive hallucination signature, the fitted AR(1) coefficients primarily reflect the global cool-down pace dictated by the diffusion schedule. Temporal persistence is therefore an intrinsic property of the sampler, confirming our core thesis: the discriminative signal for hallucination resides fundamentally in cross-token heterogeneity rather than temporal correlation.

3.5 Features and classifier

The complete denoising trace (H, ℓ, m) is distilled into **37** scalar closed-form features across five families (fully tabulated in Appendix C): *level* (means of entropy and log-probabilities, including

unmasked and non-padded variants), *dispersion* (temporal and cross-token variances), *shape of* $V(t)$ (area, peak amplitude, argmax, skewness, kurtosis, curvature), *dynamics* (finite-difference rates), and *Markovian* features (log-linear decay slopes $\alpha \approx -1/\tau_d$, offsets β, and parametric fits for C, τ_{eff}, m). We standardize all features and feed them to an L_1/L_2-regularized logistic regression, the sole learned module in our framework. Inverse regularization strength $C_{\text{reg}} \in \{10^{-3}, \ldots, 10^{2}\}$ and penalty norm are selected via stratified 5-fold cross-validation optimizing ROC-AUC (SAGA solver). Strong L_1 regularization naturally performs parameter shrinkage over collinear variance descriptors, eliminating heuristic feature selection while preserving interpretability.

4 Experimental setup

Models and generation. We evaluate LLaDA-8B-Instruct, Dream-7B-Instruct, and DiffusionGemma (26B MoE, 3.8B active) using the open-source `dLLM` library (Zhou et al., 2025) under a unified sampler interface. Because DiffusionGemma was natively unsupported, we contributed a dedicated pipeline to log its complete denoising trajectory. For comparability across models, DiffusionGemma is evaluated with a deterministic low-confidence sampler rather than its native entropy-bounded sampler, setting canvas length equal to generation length to prevent truncation. All primary generations are deterministic ($T = 0$, low-confidence remasking: at each step, the D/T most confident masked positions are fixed). We benchmark five sequence-length and step configurations $(D, T) \in \{(32, 16), (32, 32), (64, 32), (64, 64), (256, 128)\}$ on TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018), and Natural Questions (Kwiatkowski et al., 2019) ($\sim$2,500 questions each, fixed seed 42). Jobs are executed on two GPUs per run across an eight-GPU RTX 6000 Ada cluster. Main results focus on the $(32, 16)$ configuration; all remaining configurations are reported in Appendix D with identical empirical conclusions.

Labels and judge validation. Generated answer correctness was judged by Qwen3.5-9B (Qwen Team, 2026) (hosted locally via vLLM at $T = 0$) conditioned on the question, reference answer (with aliases), and model generation, applying a verbatim-match first, followed by semantic-equivalence tolerance. Empty outputs are labeled as hallucinations, while ambiguous judgments are excluded. To ensure the reliability of this automated supervision, we conducted a two-tiered validation. First, Claude Opus 5 re-annotated a stratified random sample of 504 generations (56 per model–dataset pair) blind to Qwen labels, yielding an agreement of Cohen's $\kappa = 0.93$. Second, we benchmarked the judge against 1,191 expert human validation votes across 343 unique TriviaQA question–generation pairs (evaluated across 12 sessions by computer science researchers, with 2 to 5 annotations per pair). Qwen3.5-9B matches human consensus at 95.3% agreement (327/343) and Cohen's $\kappa = 0.90$, in line with standard human–human inter-annotator agreement for factual QA (Zheng et al., 2023; Adlakha et al., 2024).

Baselines and sampling asymmetry. Perplexity and LN-Entropy are evaluated on the exact same single deterministic trajectory as DRIFTY ($N_{\text{gen}} = 1$, $T = 0$). Semantic Entropy and Lexical Similarity require $N = 5$ stochastic samples per prompt ($T = 0.5$); Semantic Entropy evaluates pairwise bidirectional entailment using DeBERTa-v2-xlarge-MNLI (He et al., 2021), while Lexical Similarity computes ROUGE-L overlaps across alternatives. Supervised methods (Lexical Similarity, TraceDet, and DRIFTY) share an identical 80/20 train/test split protocol (TraceDet with early stopping on a stratified 10% validation subset; see Appendix D).

Metrics and imbalance. Because hallucination prevalence varies substantially across models and benchmarks (Table 2), we report ROC-AUC and PR-AUC jointly. While ROC-AUC provides a threshold-independent baseline centered at 0.5, PR-AUC evaluates performance against dataset prevalence, penalizing false positives in high-risk operational regimes.

5 Results

5.1 DRIFTY dominates the AR baselines everywhere

Table 3 reports the test-split ROC-AUC and PR-AUC across all nine model–dataset pairings. DRIFTY outperforms every autoregressive baseline across all settings and metrics, while matching

Hallucination rate (32 tok. / 16 steps)	**TriviaQA**	**Natural Questions**	**HotpotQA**
LLaDA-8B	58.3%	69.4%	71.3%
Dream-7B	47.6%	56.8%	69.6%
DiffusionGemma	30.4%	56.0%	61.6%

Table 2: Hallucination prevalence per model and dataset (2,500 questions each, seed 42, $T = 0$; ambiguous judgments excluded).

		TriviaQA		Natural Questions		HotpotQA		**HW**	**s / 100**
Model	**Method**	ROC	PR	ROC	PR	ROC	PR		
Dream	Perplexity	0.507	0.481	0.604	0.693	0.542	0.703	CPU	0.21
	LN-Entropy	0.596	0.600	0.649	0.707	0.578	0.761	CPU	0.17
	Semantic Entropy	0.760	0.683	0.665	0.711	0.752	0.836	2 GPU	188
	Lexical Similarity	0.734	0.675	0.605	0.651	0.689	0.815	2 GPU	200
	TraceDet	0.830	0.810	0.701	0.779	0.774	0.871	GPU	37.63
	DRIFTY (ours)	**0.852**	**0.842**	**0.810**	**0.872**	**0.822**	**0.900**	CPU	**0.49**
LLaDA	Perplexity	0.688	0.753	0.694	0.819	0.695	0.814	CPU	0.21
	LN-Entropy	0.643	0.689	0.622	0.774	0.668	0.783	CPU	0.18
	Semantic Entropy	0.805	0.818	0.772	0.840	0.784	0.866	2 GPU	200
	Lexical Similarity	0.737	0.753	0.693	0.790	0.715	0.832	2 GPU	234
	TraceDet	0.851	0.865	0.752	0.853	0.781	0.888	GPU	34.71
	DRIFTY (ours)	**0.857**	**0.892**	**0.815**	**0.903**	**0.801**	**0.905**	CPU	**0.45**
DiffGemma	Perplexity	0.623	0.558	0.631	0.668	0.627	0.734	CPU	0.13
	LN-Entropy	0.688	0.606	0.670	0.700	0.665	0.762	CPU	0.10
	Semantic Entropy	0.801	0.607	0.686	0.661	0.681	0.720	2 GPU	459
	Lexical Similarity	0.635	0.436	0.650	0.628	0.647	0.708	2 GPU	850
	TraceDet	0.807	0.694	**0.738**	0.765	**0.779**	**0.818**	GPU	34.46
	DRIFTY (ours)	**0.833**	**0.742**	0.737	**0.778**	0.757	0.817	CPU	**0.43**

Table 3: Test-split ROC-AUC / PR-AUC, 32 tokens / 16 steps. Best per model and column in bold. TraceDet was run with original hyperparameters. Time: detector cost per 100 samples (dataset average; CPU-only for ours vs. 5 stochastic generations on 2 GPUs for multi-sample baselines, wall-clock, $\times 2$ in GPU·s). Note that DiffGemma GPU times are not directly comparable (a 2-GPU sharding forces smaller batches). All other configurations results are displayed in Appendix Section D with the same conclusions.

or exceeding the diffusion-specific TraceDet in 15 out of 18 evaluation cells. On LLaDA evaluated on TriviaQA, DRIFTY reaches 0.857 ROC-AUC, improving over Semantic Entropy (0.805), TraceDet (0.851), and Perplexity (0.688, an advantage of nearly 17 points).

These performance gains are even more pronounced under PR-AUC. For example, on Dream evaluated over Natural Questions, DRIFTY scores 0.872 compared to 0.711 for Semantic Entropy ($+16$ points). The few instances where DRIFTY trails TraceDet are marginal (deficits ≤ 2.2 points) and occur exclusively on DiffusionGemma, where uniform-state re-noising disrupts monotonic entropy decay. Standard AR baselines degrade because evaluating predictive distributions strictly at token commitment discards crucial temporal context, whereas cross-token trace dispersion retains the full dynamics of model hesitation.

Detector latency and compute cost. Because DRIFTY operates directly on the trace of a single deterministic generation ($N_{\text{gen}} = 1$), feature extraction and linear classification require only 0.43–0.49 s of CPU time per 100 samples. In contrast, multi-sample baselines require 5 stochastic generations followed by NLI cross-encoding or pairwise lexical alignments, taking between 188 s and 850 s on two GPUs. Furthermore, DRIFTY matches or surpasses TraceDet (0.80–0.86 ROC-AUC on LLaDA and Dream) without training or evaluating a neural Transformer detector, reducing evaluation latency by roughly two orders of magnitude.

5.2 CROSS-DATASET TRANSFER AND ARCHITECTURAL BOUNDARIES

Task generalization. To verify that extracted dispersion signatures reflect general hallucination dynamics rather than dataset-specific artifacts, we evaluate out-of-domain transfer across TriviaQA, Natural Questions, and HotpotQA (complete transfer matrices in Appendix E). Detectors trained on one task transfer across domains without collapse: for instance, LLaDA trained on TriviaQA retains 0.793 ROC-AUC when evaluated on HotpotQA (matching in-domain performance within 0.01 points), demonstrating that trace heterogeneity captures shared hesitation mechanisms across both direct retrieval and multi-hop reasoning.

Uniform-state diffusion stress test: DiffusionGemma. Hallucination detection in DiffusionGemma proves systematically more challenging across all evaluated methods: DRIFTY attains 0.833 ROC-AUC on TriviaQA and approximately 0.75 on the remaining benchmarks, compared to 0.80–0.86 on LLaDA and Dream. We attribute this attenuation to two factors. Mechanistically, DiffusionGemma employs uniform-state diffusion with iterative re-noising, violating the monotonic decay assumption $\varepsilon(t) \approx Ce^{-t/\tau}$ characteristic of absorbing-state models and dampening log-linear slope signals. Statistically, DiffusionGemma exhibits a lower base hallucination rate (30.4% on TriviaQA), shifting evaluation toward rare-event classification where precision metrics are naturally more sensitive.

6 LIMITATIONS

While effective and computationally light, our approach presents several operational boundaries:

Scope of generation and sequence length. Our evaluation focuses on closed-book question answering with short sequence lengths ($D \leq 256$) without extended chain-of-thought reasoning. However, because d-LLMs predominantly generate long-form documents block-by-block (semi-autoregressively), trace dispersion can be naturally computed at the block level. Furthermore, because our key heterogeneity metrics explicitly condition on still-masked tokens ($m_{t,d} = 1$), they are structurally insulated from dilution by trailing padding tokens in variable-length completions.

Deterministic decoding ($T = 0$). We intentionally restrict generations to greedy decoding ($T = 0$) to ensure exact reproducibility and a strictly aligned unmasking schedule π for transposing AR baselines. Evaluating trace dynamics under stochastic sampling ($T > 0$) introduces exploratory entropy fluctuations that warrant future investigation.

Trace storage overhead. Operating post hoc requires buffering intermediate distributions (H, ℓ, m) during generation, incurring a higher memory footprint than autoregressive single-token logging, though feature extraction remains sub-second on CPU.

Supervision and persistent errors. DRIFTY requires labeled validation trajectories to fit its regularized logistic head, inheriting an estimated $\sim 5\%$ label noise margin from the LLM judge. Moreover, because dispersion fundamentally tracks hesitation, pathological failures where a model commits to a wrong entity at step $t = 0$ with uniformly high confidence remain an intrinsic blind spot shared by trajectory-based uncertainty metrics.

7 CONCLUSION

We asked what constitutes a sufficient low-dimensional summary of the denoising trace for hallucination detection. Our findings demonstrate that while average confidence carries useful signal, cross-token inhomogeneity carries far more. A Markovian convergence model formalizes this dynamic, showing that disparate token difficulty produces the characteristic delayed, asymmetric variance peak of hallucinations, while stochastic Ornstein–Uhlenbeck modeling rules out temporal persistence as an alternative explanation. Distilled into 37 closed-form features and regularized logistic regression, DRIFTY dominates AR baselines across all evaluated settings, matches or exceeds neural detectors at CPU speed, and transfers reliably across tasks. Ultimately, hesitation—the spatio-temporal texture of uncertainty as answers take shape—provides a lightweight and principled foundation for uncertainty quantification in diffusion language models.

REPRODUCIBILITY STATEMENT

To ensure full reproducibility of our empirical results, all primary text generations were conducted deterministically (decoding temperature $T = 0$, fixed random seed 42, and low-confidence remasking), with non-zero temperature restricted exclusively to multi-sample baselines requiring stochastic completions. Diffusion generations were executed using the open-source `dLLM` library (Zhou et al., 2025). All benchmark datasets were retrieved via HuggingFace through the `datasets` library (Lhoest et al., 2021). Annotation labels were produced according to the LLM-as-a-judge protocol detailed in Section 4, with inference hosted locally using `vLLM` (Kwon et al., 2023). Feature extraction and classifier training rely on standard scientific Python libraries (NumPy and scikit-learn for regularized logistic regression and stratified 5-fold cross-validation). Our benchmark runs of TraceDet use official repository defaults without hyperparameter tuning on identical splits (Appendix D). Our complete, anonymized codebase—encompassing trace logging, closed-form feature extractors, judge prompt templates, and evaluation scripts—is publicly accessible at `https://anonymous.4open.science/r/drifty-3762/`.

ETHICS STATEMENT

This work aims to enhance the factual reliability of diffusion language models by detecting hallucinations; it does not introduce harmful applications or handle sensitive personal data. Human validation described in this paper was limited to voluntary expert evaluation of factual question–answer pairs from standard public benchmarks. Because miscalibrated detection could lead to false positives (improperly filtering or suppressing accurate generations), we evaluate models across complete ROC and PR threshold curves rather than mandating a single rigid decision threshold. Given an estimated $\sim 5\%$ label noise margin in automated judges, we advise that high-stakes applications pair automated trace scores with human-in-the-loop oversight.

AI USE STATEMENT

Generative AI assistants (Claude Opus 5 and GLM-5.3 (GLM-5-Team et al., 2026)) were employed to assist in drafting prose, editing style, and scaffolding auxiliary utility scripts. All experimental protocols, mathematical formulations, software implementations, and analyses were conceived, verified, and finalized by the authors. No generative AI tools were used to produce experimental data or benchmark results, with the sole exception of the automated LLM Judge labels described and human-validated in Section 4.

REFERENCES

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699, 2024.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 17981–17993, 2021. URL `https://arxiv.org/abs/2107.03006`.

Heli Ben-Hamu, Itai Gat, Daniel Severo, Niklas S Nolte, and Brian Karrer. Accelerated sampling from masked diffusion models via entropy bounded unmasking. *Advances in Neural Information Processing Systems*, 38:55981–56007, 2026.

Shenxu Chang, Junchi Yu, Weixing Wang, Yongqiang Chen, Jialin Yu, Philip Torr, and Jindong Gu. Tracedet: Hallucination detection from the decoding trace of diffusion large language models. *arXiv preprint arXiv:2510.01274*, 2025. doi: 10.48550/arxiv.2510.01274.

DiffusionGemma Team, Adrien Ali Taïga, James Assiene, Daniele Calandriello, Rahma Chaabouni, João Gante, Tamara von Glehn, Nate Keating, Chris Knutsen, Martin Kukla, et al. Diffusiongemma technical report. *arXiv preprint arXiv:2608.00146*, 2026.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 628(8009):1–7, 2024. doi: 10.1038/s41586-024-07421-0.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 539–555, 2020.

Crispin Gardiner. *Stochastic methods*, volume 4. Springer Berlin Heidelberg, 2009.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2006.03654.

Arshia Hemmat, Philip Torr, Yongqiang Chen, and Junchi Yu. Tdgnet: Hallucination detection in diffusion language models via temporal dynamic graphs. *arXiv preprint arXiv:2602.08048*, 2026.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in neural information processing systems*, 34:12454–12465, 2021.

Fred Jelinek, Robert L. Mercer, Lalit R. Bahl, and James K. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1): S63–S63, 1977. doi: 10.1121/1.2016299.

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

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa as a resource for open domain question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariants for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023. URL https://arxiv.org/abs/2302.09664.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 175–184, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-demo.21.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6107–6117, 2025.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2002.07650.

Shen Nie, Fengqi Zhu, Chao Du, Tianyu Pang, Qian Liu, Guangtao Zeng, Min Lin, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2024. URL https://arxiv.org/abs/2406.07524.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Artem Vazhentsev, Vladislav Smirnov, David Li, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. Uncertainty quantification for large language diffusion models, 2026. URL https://arxiv.org/abs/2605.14570.

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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Zhanhui Zhou, Lingjie Chen, Hanghang Tong, and Dawn Song. dllm: Simple diffusion language modeling. https://github.com/ZHZisZZ/dllm, 2025.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. LLaDA 1.5: Variance-reduced preference optimization for large language diffusion 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)*, pp. 11425–11460, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.524. URL https://aclanthology.org/2026.acl-long.524/.

This appendix accompanies our main paper *Heterogeneous Uncertainty Along Denoising Traces Signals Hallucinations in Diffusion Language Models*. It provides full mathematical derivations, statistical details of the Ornstein–Uhlenbeck and AR(1) analyses, the exhaustive dictionary of trace features, and comprehensive evaluation tables across all token and step configurations.

A Derivation of the entropy decay

Under the Markovian convergence model introduced in Section 3.3, the predictive distribution at position d assigns probability mass $1-\varepsilon$ to the target token i_d^* and distributes the remaining mass $\varepsilon/(|\mathcal{V}|-1)$ uniformly across all other vocabulary entries, with $\varepsilon(t) = Ce^{-t/\tau}$. The exact Shannon entropy is:

$$\begin{aligned} S(t) &= -(1-\varepsilon)\ln(1-\varepsilon) - \sum_{j\neq i^*} \frac{\varepsilon}{|\mathcal{V}|-1} \ln \frac{\varepsilon}{|\mathcal{V}|-1} \\ &= -(1-\varepsilon)\ln(1-\varepsilon) - \varepsilon \ln \varepsilon + \varepsilon \ln(|\mathcal{V}|-1). \end{aligned} \tag{7}$$

In the regime where the model solidifies its prediction ($\varepsilon \ll 1$), we apply the first-order Taylor expansion $\ln(1-\varepsilon) \approx -\varepsilon$, yielding:

$$\begin{aligned} S(t) &\approx \varepsilon\big(1 - \ln \varepsilon + \ln(|\mathcal{V}|-1)\big) \\ &= C\, e^{-t/\tau} \left(\frac{t}{\tau} + \ln \frac{|\mathcal{V}|-1}{C} + 1 \right). \end{aligned} \tag{8}$$

B AR(1) analysis details

For each class (correct vs. hallucinated) and each spatio-temporal coordinate (t, d), the AR(1) process defined in main text Equation (6) is fitted via ordinary least squares across all sequences within the class, pooling samples at token position d. Because each regression is univariate ($p = 1 \ll n$), the estimator is statistically well-conditioned and robust to sample noise.

Figure 5 compares the mean and variance of the masked entropy trajectories generated by the two fitted AR(1) models against empirical generation traces. The synthetic trajectories produced by the class-specific models are virtually superposable, confirming that AR(1) dynamics reflect the global cool-down pace of the diffusion schedule rather than a label-specific hallucination signature (Section 3.4).

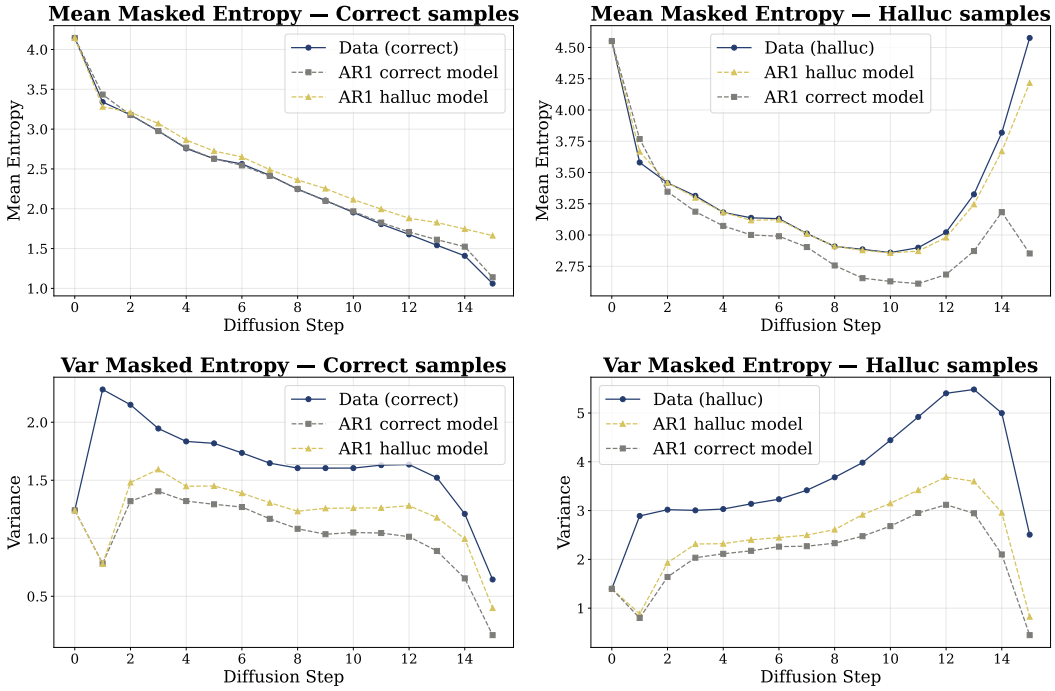

Figure 5: Masked entropy mean and variance: empirical trajectories versus synthetic paths sampled from class-conditional AR(1) models fitted on correct and hallucinated samples. The fitted dynamics coincide across classes, demonstrating that temporal persistence does not provide discriminative signal.

C Full feature list and correlation analysis

Figure 6 provides dataset-wide empirical support for our feature design on Dream (32 tokens/16 steps). The top five predictors by absolute Pearson correlation $|r|$ are all cross-token dispersion metrics: `VarMaskedEntropyAcrossTokens` (0.531), `AUCVarMaskedEntropyCurve` (0.530), `VarLogProbMaskedAcrossTokens` (0.477), `MaxVarMaskedEntropyCurve` (0.468), and `ArgmaxVarMaskedEntropyCurve` (0.452). In contrast, the strongest level-based metric (`MeanMaskedEntropy`) achieves only $r = 0.306$. The most negative correlation is observed for `MeanLogProbMaskedNoPad` (-0.350), confirming that sustained confidence over non-padded positions signals factual accuracy. Because top dispersion features represent correlated descriptors of a single underlying phenomenon, regularized logistic regression (L_1/L_2) naturally performs parameter shrinkage over these collinear variants. Ablations across feature families (Baseline, Markov, and DRIFTY) across all generation lengths are detailed in Appendix D.

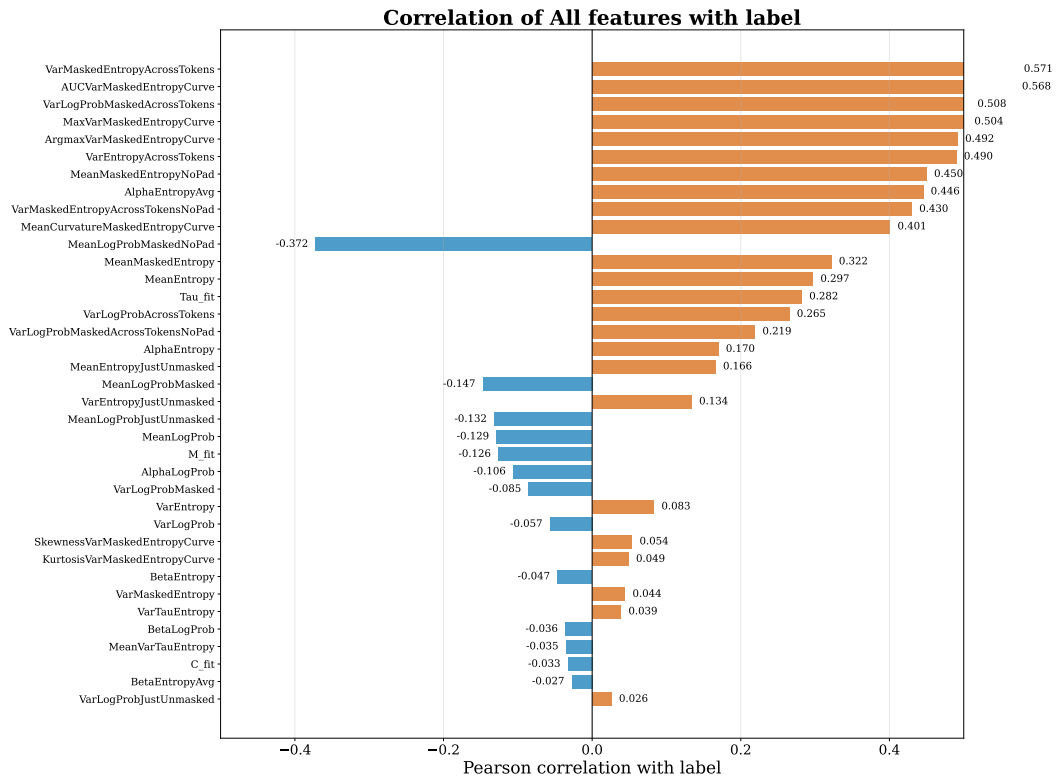

Figure 6: Pearson correlation r between each of the 37 closed-form trace features and the binary hallucination label (evaluated on Dream, 16 steps, 32 tokens). Cross-token dispersion statistics dominate the top ranks.

C.1 Trace statistics feature dictionary

Table 4 details the 37 closed-form scalar features extracted from the denoising trace (H, ℓ, m), categorized across their five structural families.

Family	Features	What they measure
Level	`MeanEntropy`, `MeanMaskedEntropy`, `MeanEntropyJustUnmasked` + log-probability analogues + `NoPad` variants	Average predictive confidence along the trajectory
Dispersion	`VarEntropy`, `VarEntropyAcrossTokens`, `VarMaskedEntropy`, `VarMaskedEntropyAcrossTokens`, `VarEntropyJustUnmasked` + log-probability analogues + `NoPad` variants	Temporal and cross-token uncertainty dispersion
Dynamics	`MeanVarTauEntropy`, `VarTauEntropy`	Finite-difference rates of entropy change
Shape of $V(t)$	`AUC`, `Max`, `Argmax`, `Skewness`, `Kurtosis`, `MeanCurvature` of the variance curve	Amplitude, timing, and asymmetry of the heterogeneity peak
Markovian	`AlphaEntropyAvg/Var`, `BetaEntropyAvg` (log-linear fits of $\ln H_{t,d}$ vs. t), `C_fit`, `Tau_fit`, `M_fit` (parametric fits of Equation (4) on hardest tokens)	Token-level convergence rates and sequence-level decay scale

Table 4: The 37 closed-form features extracted from the denoising trace.

C.2 AR(1) PROCESS PARAMETERS

Table 5 summarizes the stochastic autoregressive parameters estimated during the temporal persistence analysis in Section 3.4.

Family	Parameters	What they measure
AR(1)	φ, c, σ^{ε} per spatio-temporal coordinate	Local autoregressive persistence of entropy fluctuations (evaluated in Section 3.4; not included in the 37 DRIFTY classifier features)

Table 5: Estimated parameters for the AR(1) Ornstein–Uhlenbeck discretization.

D ADDITIONAL EXPERIMENTAL RESULTS ACROSS CONFIGURATIONS

All tables in this section report test-split ROC-AUC and PR-AUC evaluated against the validated judge labels from Section 4. Baseline (level, dispersion, and shape features), Markov (convergence rates only), and DRIFTY (combined set) represent our detector variants. TraceDet reproduction follows the protocol detailed below.

TraceDet reproduction details. We executed TraceDet from its official public repository using default settings from `train_main.py`: a 3-layer, 1-head Transformer with feed-forward dimension 8, temporal positional encoding dimension $d_{\text{pe}} = 8$, and dropout 0.2. Optimization employed AdamW (learning rate $2 \cdot 10^{-4}$, weight decay 0.1, batch size 64) for up to 100 epochs, incorporating early stopping (patience 20) on validation ROC-AUC (stratified 10% split of the training set) and `ReduceLROnPlateau` scheduling after epoch 20, with mask penalty $\beta = 0.1$ and seed 42. Training and evaluation were conducted on identical splits using two RTX 6000 Ada GPUs per job.

D.1 64 TOKENS / 32 STEPS

Model	Method	TriviaQA ROC	TriviaQA PR	N. Questions ROC	N. Questions PR	HotpotQA ROC	HotpotQA PR	Time /100 (s)
Dream	Perplexity	0.515	0.491	0.633	0.692	0.532	0.685	0.31
	LN-Entropy	0.626	0.655	0.629	0.699	0.592	0.764	0.25
	Semantic Entropy	0.738	0.687	0.657	0.688	0.757	0.813	188
	Lexical Similarity	0.722	0.680	0.618	0.660	0.668	0.777	202
	TraceDet	0.869	**0.867**	0.739	0.808	0.740	0.824	—
	DRIFTY (ours)	**0.874**	0.856	**0.771**	**0.819**	**0.826**	**0.884**	**0.68**
LLaDA	Perplexity	0.547	0.625	0.591	0.769	0.583	0.766	0.34
	LN-Entropy	0.511	0.605	0.589	0.768	0.570	0.762	0.28
	Semantic Entropy	0.761	0.779	0.763	0.849	0.727	0.832	198
	Lexical Similarity	0.674	0.698	0.682	0.800	0.715	0.849	230
	TraceDet	0.812	0.832	0.757	**0.879**	0.777	0.884	—
	DRIFTY (ours)	**0.851**	**0.873**	**0.800**	**0.879**	**0.814**	**0.911**	**0.68**
DiffGemma	Perplexity	0.672	0.524	0.575	0.606	0.590	0.711	0.39
	LN-Entropy	0.763	0.611	0.643	0.658	0.667	0.769	0.33
	Semantic Entropy	0.776	0.588	0.698	0.665	0.724	0.774	454
	Lexical Similarity	0.630	0.421	0.625	0.617	0.634	0.719	881
	TraceDet	0.812	0.668	0.733	0.745	**0.742**	**0.812**	—
	DRIFTY (ours)	**0.841**	**0.722**	**0.745**	**0.754**	**0.742**	0.808	**0.72**

Table 6: Test-split ROC-AUC / PR-AUC, 64 tokens / 32 steps. Best per model and column among displayed methods in bold. TraceDet from our runs (original hyperparameters, untuned). Same protocol as main Table 1.

D.2 32 TOKENS / 32 STEPS

		TriviaQA		Natural Questions		HotpotQA	
Model	**Method**	ROC	PR	ROC	PR	ROC	PR
Dream	Perplexity	0.583	0.572	0.652	0.741	0.555	0.709
	LN-Entropy	0.660	0.679	0.670	0.780	0.575	0.753
	Semantic Entropy	0.740	0.690	0.654	0.736	0.775	0.838
	Lexical Similarity	0.704	0.654	0.573	0.695	0.702	0.816
	TraceDet	**0.862**	0.850	0.741	0.851	0.815	0.898
	Baseline	0.849	0.842	**0.786**	**0.865**	0.840	0.893
	Markov	0.844	0.809	0.745	0.829	0.800	0.865
	DRIFTY	**0.862**	**0.854**	0.781	0.859	**0.860**	**0.913**
LLaDA	Perplexity	0.550	0.586	0.620	0.770	0.553	0.679
	LN-Entropy	0.557	0.582	0.590	0.754	0.545	0.671
	Semantic Entropy	0.759	0.757	0.759	0.838	0.748	0.812
	Lexical Similarity	0.697	0.704	0.645	0.751	0.669	0.769
	TraceDet	0.814	0.810	0.761	0.875	0.799	0.884
	Baseline	0.825	0.840	**0.834**	**0.911**	**0.834**	**0.904**
	Markov	0.799	0.809	0.753	0.853	0.774	0.845
	DRIFTY	**0.836**	**0.847**	0.832	0.904	0.833	**0.904**
DiffGemma	Perplexity	0.572	0.465	0.534	0.569	0.563	0.676
	LN-Entropy	0.642	0.539	0.558	0.601	0.602	0.708
	Semantic Entropy	0.754	0.558	0.693	0.651	0.716	0.749
	Lexical Similarity	0.641	0.435	0.659	0.634	0.653	0.703
	TraceDet	**0.820**	**0.715**	0.720	0.726	**0.756**	**0.818**
	Baseline	0.808	0.700	0.723	0.752	0.745	0.802
	Markov	0.788	0.672	0.702	0.710	0.735	0.804
	DRIFTY	0.818	0.710	**0.729**	**0.758**	0.748	0.812

Table 7: Test-split ROC-AUC / PR-AUC, 32 tokens / 32 steps. Best per model and column in bold.

D.3 64 TOKENS / 64 STEPS

		TriviaQA		Natural Questions		HotpotQA	
Model	**Method**	ROC	PR	ROC	PR	ROC	PR
Dream	Perplexity	0.528	0.510	0.601	0.703	0.545	0.723
	LN-Entropy	0.594	0.572	0.615	0.723	0.579	0.750
	Semantic Entropy	0.762	0.700	0.683	0.767	0.759	0.842
	Lexical Similarity	0.722	0.654	0.625	0.739	0.657	0.776
	TraceDet	0.865	0.849	0.738	0.835	0.824	0.895
	Baseline	0.853	0.823	0.771	0.858	0.836	0.914
	Markov	0.818	0.761	0.747	0.833	0.788	0.866
	DRIFTY	**0.869**	**0.855**	**0.781**	**0.859**	**0.840**	**0.918**
LLaDA	Perplexity	0.540	0.614	0.588	0.748	0.569	0.738
	LN-Entropy	0.525	0.604	0.583	0.746	0.564	0.735
	Semantic Entropy	0.755	0.771	0.756	0.842	0.743	0.838
	Lexical Similarity	0.708	0.746	0.700	0.804	0.684	0.807
	TraceDet	0.827	0.863	0.803	0.905	0.749	0.844
	Baseline	0.830	0.872	0.823	0.915	0.809	0.901
	Markov	0.822	0.863	0.820	0.907	0.767	0.876
	DRIFTY	**0.849**	**0.884**	**0.839**	**0.920**	**0.814**	**0.904**
DiffGemma	Perplexity	0.532	0.393	0.575	0.574	0.591	0.714
	LN-Entropy	0.640	0.509	0.642	0.643	0.655	0.750
	Semantic Entropy	0.774	0.561	0.676	0.638	0.687	0.740
	Lexical Similarity	0.624	0.404	0.621	0.613	0.629	0.708
	TraceDet	0.803	0.660	0.702	0.705	0.744	**0.811**
	Baseline	0.834	0.725	0.723	0.714	0.746	0.785
	Markov	0.819	0.698	**0.735**	0.713	0.719	0.781
	DRIFTY	**0.853**	**0.756**	**0.735**	**0.739**	**0.749**	0.788

Table 8: Test-split ROC-AUC / PR-AUC, 64 tokens / 64 steps. Best per model and column in bold.

D.4 256 TOKENS / 128 STEPS

Model	Method	TriviaQA ROC	TriviaQA PR	Natural Questions ROC	Natural Questions PR	HotpotQA ROC	HotpotQA PR
Dream	Perplexity	0.508	0.460	0.498	0.664	0.529	0.678
	LN-Entropy	0.531	0.477	0.501	0.663	0.549	0.695
	Semantic Entropy	0.769	0.715	0.675	0.781	0.766	0.834
	Lexical Similarity	0.753	0.723	0.587	0.737	0.662	0.764
	TraceDet	0.832	0.785	0.697	0.818	**0.810**	**0.897**
	Baseline	0.830	**0.837**	0.733	0.848	0.805	0.879
	Markov	0.794	0.744	0.642	0.781	0.732	0.805
	DRIFTY	**0.843**	0.834	**0.761**	**0.873**	0.805	0.882
LLaDA	Perplexity	0.533	0.603	0.381	0.627	0.461	0.689
	LN-Entropy	0.522	0.599	0.371	0.630	0.443	0.679
	Semantic Entropy	0.755	0.771	0.733	0.840	0.723	0.836
	Lexical Similarity	0.691	0.733	0.631	0.771	0.671	0.802
	TraceDet	**0.870**	**0.906**	**0.835**	**0.918**	0.864	0.941
	Baseline	0.851	0.889	0.787	0.898	0.867	0.941
	Markov	0.813	0.863	0.777	0.897	0.838	0.918
	DRIFTY	0.857	0.897	0.808	0.908	**0.870**	**0.944**
DiffGemma	Perplexity	0.571	0.420	0.480	0.487	0.533	0.672
	LN-Entropy	0.624	0.450	0.482	0.483	0.548	0.678
	Semantic Entropy	0.719	0.474	0.609	0.569	0.647	0.729
	Lexical Similarity	0.645	0.416	0.504	0.496	0.609	0.721
	TraceDet	0.738	0.557	0.709	0.721	0.690	0.789
	Baseline	0.784	**0.677**	0.690	0.692	0.690	0.797
	Markov	0.781	0.615	0.681	0.696	0.672	0.771
	DRIFTY	**0.818**	0.674	**0.713**	**0.727**	**0.698**	**0.803**

Table 9: Test-split ROC-AUC / PR-AUC, 256 tokens / 128 steps. Best per model and column in bold.

E CROSS-DATASET GENERALIZATION

Model	Train	TriviaQA ROC	TriviaQA PR	NQ ROC	NQ PR	HotpotQA ROC	HotpotQA PR
LLaDA	TriviaQA	**0.852**	**0.889**	0.744	0.892	0.793	0.914
	NQ	0.812	0.862	**0.779**	**0.908**	0.764	0.899
	HotpotQA	0.838	0.877	0.752	0.897	**0.807**	**0.919**
Dream	TriviaQA	**0.859**	**0.848**	0.731	0.840	0.799	0.897
	NQ	0.797	0.804	**0.778**	**0.869**	0.770	0.886
	HotpotQA	0.835	0.825	0.750	0.851	**0.818**	**0.905**
DiffGemma	TriviaQA	**0.820**	**0.699**	0.726	0.786	0.751	0.824
	NQ	0.796	0.644	**0.748**	**0.804**	0.751	0.824
	HotpotQA	0.802	0.662	0.736	0.791	**0.763**	**0.831**

Table 10: Cross-dataset transfer (32 tokens / 16 steps). Rows: train set; columns: test set; diagonal bold. Off-diagonal PR-AUC sometimes exceeds the diagonal.

Table 10 shows transfer without collapse. LLaDA trained on TriviaQA and tested on HotpotQA keeps 0.793 ROC vs. 0.807 direct (gap < 2 points); Dream and DiffGemma behave alike. Off-diagonal PR-AUC can even exceed the diagonal. Trace heterogeneity thus captures a shared mechanism across factual, multi-hop and search-query reasoning, not a dataset quirk.