

OSCAR: Orchestrated Self-verification and Cross-path Refinement

Yash Shah*

Arizona State University
yshah124@asu.edu

Abhijit Chakraborty*

Arizona State University
achakr40@asu.edu

Naresh Kumar Devulapally

University at Buffalo, SUNY
devulapa@buffalo.edu

Vishnu Suresh Lokhande

University at Buffalo, SUNY
vishnulo@buffalo.edu

Vivek Gupta

Arizona State University
vgupt140@asu.edu

Abstract

Diffusion language models (DLMs) expose their denoising trajectories, offering a natural handle for inference-time control; accordingly, an ideal hallucination mitigation framework should intervene during generation using this model-native signal rather than relying on an externally trained hallucination classifier. Toward this, we formulate *commitment uncertainty localization*: given a denoising trajectory, identify token positions whose cross-chain entropy exceeds an unsupervised threshold before factually unreliable commitments propagate into self-consistent but incorrect outputs. We introduce a suite of trajectory-level assessments, including a cross-chain divergence-at-hallucination (CDH) metric, for principled comparison of localization methods. We also introduce OSCAR, a training-free inference-time framework operationalizing this formulation. OSCAR runs N parallel denoising chains with randomized reveal orders, computes cross-chain Shannon entropy to detect high-uncertainty positions, and then performs targeted remasking conditioned on retrieved evidence. Ablations confirm that localization and correction contribute complementary gains, robust across $N \in \{4, 8, 16\}$. On TriviaQA, HotpotQA, RAGTruth, and CommonsenseQA using LLaDA-8B and Dream-7B, OSCAR enhances generation quality by significantly reducing hallucinated content and improving factual accuracy through uncertainty-guided remasking, which also facilitates more effective integration of retrieved evidence. Under a unified LLM-as-Judge protocol, its native entropy-based uncertainty signal *matches* that of the strongest specialized trained detectors, without requiring any supervision or labeled data, while additionally providing a targeted correction stage that no detector baseline offers thus highlighting an inherent capacity of diffusion language models to identify factual uncertainty that is not present in the sequential token commitment structure of autoregressive models. We release the codebase¹ to support future research on localization and uncertainty-aware generation in DLMs.

1 Introduction

The Transformer architecture (Vaswani et al., 2017) revolutionized sequence modeling by replacing recurrence with self-attention, leading to autoregressive (AR) decoding for token generation. This enables hallucination detection tools like perplexity thresholding and retrieval-augmented verification (Quevedo et al., 2025; Malinin & Gales, 2021; Lewis et al., 2020; Tonmoy et al., 2024). Diffusion language models (DLMs) generate text through iterative demasking from masked sequences (Nie et al., 2025; Ye et al., 2025). Unlike autoregressive

*Equal contribution.

¹<https://github.com/coral-lab-asu/Oscar-DLLM-Hallucination-Reduction>

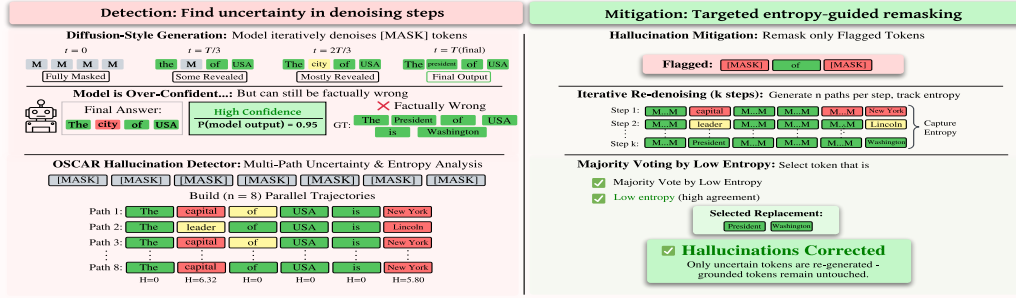


Figure 1: Schematic overview of the OSCAR pipeline: detection of commitment uncertainty in the denoising steps, followed by targeted entropy-guided remasking.

LLMs, DLMs maintain a trajectory (Li et al., 2025) of intermediate states encoding uncertainty, impacting question answering where early errors cause hallucination crystallization (Lu et al., 2025). Hallucination detection and DLM denoising dynamics developed largely in parallel: detection literature trains classifiers to recognize hallucinations post-generation, while DLM literature characterizes trajectory-level phenomena without exploiting them for correction. Yet DLM denoising’s order-dependence of token commitment directly yields the uncertainty signal that detection classifiers approximate (Bloem-Reddy & Teh, 2020; Chytas et al., 2024); OSCAR bridges this gap by transforming a diagnostic signal into a correction mechanism.

Prior work on hallucination detection includes *output-based* methods (Kuhn et al., 2023; Lin et al., 2023; Ren et al., 2022; Malinin & Gales, 2021) analyzing final output, *latent-based* methods (Burns et al., 2022; Park et al., 2025; Azaria & Mitchell, 2023; Chen et al., 2024) probing internal representations, and *trajectory-based* methods TraceDet (Chang et al., 2025), DynHD (Qian et al., 2026) exploiting multi-step dynamics. Trajectory-based methods achieve strongest results but require trained classifiers that may not transfer across domains. These methods only flag hallucinations without correction. This is due to AR-based tools’ incompatibility with bidirectional denoising, lack of ground-truth signals, and underutilized trajectory data. (Niu et al., 2024; Min et al., 2023; Tonmoy et al., 2024).

We argue that rather than training classifiers to recognize hallucinations after generation, we should exploit the *native uncertainty signal* from DLMs. Our key observation is:

If multiple denoising chains, each following a different randomized reveal order, disagree on a token position, that position is factually uncertain; regardless of how confident any single chain appears.

This leads to **commitment uncertainty localization** (§3): given parallel denoising trajectories, identify token positions where cross-chain Shannon entropy exceeds an unsupervised threshold, *before* premature commitments lead to incorrect outputs; an **adaptive remasking and re-denoising phase** then selectively re-generates the uncertain spans. We implement these in OSCAR (Orchestrated Selective Correction via Adaptive Remasking), a training-free inference-time framework that runs N parallel denoising chains with randomized reveal orders, computes cross-chain entropy to localize uncertainties, and performs targeted remasking with retrieved evidence at modest runtime cost. The principle that disagreement among diverse generations reveals uncertainty is shared with self-consistency, semantic entropy, and SelfCheckGPT (Wang et al., 2023; Kuhn et al., 2023; Manakul et al., 2023), and we do not claim it as novel; our specific contribution is the combination of *randomized reveal orders with cross-chain entropy localization, post-hoc span-level remasking, and targeted retrieval conditioning* into a training-free detection-and-correction pipeline exploiting the commitment-order structure of masked diffusion, to our knowledge, the first such inference-time pipeline for DLMs.

Our contributions: (a) We identify and demonstrate a *native uncertainty signal* in DLMs: cross-chain Shannon entropy computed across parallel denoising trajectories with randomized reveal orders. Under a unified LLM-as-Judge protocol, this training-free signal matches

the strongest trained trajectory detectors (statistically tied with DynHD on both models) without any external supervision, an uncertainty measure structurally absent in autoregressive models. (b) We formalize *commitment uncertainty localization* and introduce the cross-chain divergence-at-hallucination (CDH) metric, which identifies factually unreliable token positions within denoising trajectories, substantially outperforming both trained and confidence-based localizers. (c) We propose OSCAR, the first inference-time hallucination mitigation framework for DLMs, which corrects hallucinations via targeted re-masking and re-denoising of uncertain spans, yielding significant gains in generation quality and factual accuracy across benchmarks at modest computational overhead. (d) We provide a mechanistic analysis of hallucination *crystallization timing* during denoising, revealing task-dependent patterns that support early-exit detection and adaptive, task-specific intervention.

2 Related Work

Hallucination detection in autoregressive (AR) language models has developed through three main approaches. Uncertainty-based methods estimate confidence by examining output token distributions (Malinin & Gales, 2021; Ren et al., 2022). Semantic entropy methods group equivalent outputs before calculating entropy to capture meaning-level uncertainty (Kuhn et al., 2023; Quevedo et al., 2025). Probe-based methods train lightweight classifiers on internal model activations to identify hallucinations (Azaria & Mitchell, 2023; Burns et al., 2022). SelfCheckGPT evaluates agreement across multiple generations but lacks a correction mechanism (Manakul et al., 2023), and ROUGE-based evaluation has been shown inferior to LLM-as-Judge protocols (Janiak et al., 2025). None of these methods address the parallel bidirectional denoising architecture of DLMs, necessitating DLM-specific approaches.

Hallucination detection in diffusion language models (DLMs) has recently gained attention as a promising area of research. Methods such as TraceDet model the denoising process as an action trace and identify the most informative sub-traces using the Information Bottleneck principle, resulting in a 15.2% average AUROC improvement over output-based baselines. DynHD tackles the issue of information density imbalance across token positions by constructing semantic-aware evidence and introducing a reference evidence generator that learns the expected uncertainty evolution, detecting hallucinations by measuring deviations from this reference. EigenScore (Shoushtari et al., 2025) leverages the eigenvalue spectrum of the posterior covariance induced by the diffusion model as an out-of-distribution (OOD) detection signal. While these approaches effectively use the denoising trajectory as diagnostic input for classification, none modifies or controls the denoising process itself: the trajectory remains a passive signal rather than an active means for correction.

Remasking at inference time in DLMs. Revising already-committed tokens is an established inference paradigm. Mask-Predict (Ghazvininejad et al., 2019) iteratively remasks the lowest-confidence tokens in non-autoregressive translation, and Token-Critic (Lezama et al., 2022) trains a critic that revokes and resamples committed tokens in masked image generation. For masked DLMs specifically, ReMDM (Wang et al., 2025) introduces a training-free sampler with a confidence-based remasking variant demonstrated on LLaDA-8B, and RemeDi (Huang et al., 2026) trains a per-token confidence head for self-reflective remasking. All of these methods localize revision targets via *single-trajectory confidence*, whether native or learned. OSCAR differs in the localization criterion: cross-chain entropy under reveal-order diversification. Because an early incorrect commitment acts as a contextual attractor (§3.1), within-chain confidence is unreliable precisely at hallucinated positions: our confidence-based baseline (ReMDM-conf, §5.3) localizes barely above random (CDH(20%) 27.6 vs. 67.3 for cross-chain entropy) and trails OSCAR on every mitigation metric (Table 5).

Existing approaches can be viewed as restricted instantiations of a general principle: *disagreement among diverse generations reveals factual uncertainty*. We do not claim this principle as novel. Independent resampling (SelfCheckGPT (Manakul et al., 2023)) measures total output variance but conflates sampling noise with factual uncertainty because each resample traverses the full generation process independently. Trajectory classifiers (TraceDet,

DynHD) learn to recognize hallucination-correlated patterns in the denoising trace but require labeled data and provide no correction affordance. Cross-chain entropy under reveal-order diversification holds the model and sampling distribution fixed and varies *only* the token-commitment order, isolating the *commitment-order* component of disagreement directly, the precise structural source of DLM hallucination, without training; the compute-matched ablation in §6.2 quantifies this isolation empirically. OSCAR operationalizes this signal end-to-end: unlike TraceDet, DynHD, and EigenScore, which use the trajectory only for classification, it transforms the denoising trajectory from a passive diagnostic input into an active control mechanism, the first training-free, inference-time framework using the diffusion model’s own uncertainty for both detection and targeted correction.

3 OSCAR: Commitment Uncertainty Localization and Correction

We present OSCAR in four parts mirroring our contributions: the structural failure mode and its measurable, model-native uncertainty signal (§3.1; a); the localization formulation and CDH metric (§3.2; b); the detection-and-correction pipeline (§3.3; c); and the signal’s temporal structure via crystallization dynamics (§3.4; d).

3.1 A Native Uncertainty Signal in Diffusion Language Models

A DLM generates response $\mathbf{y} = (y_1, \dots, y_L)$ to query q through T denoising steps. At $t=0$ all positions are masked; at each step the model predicts $p_\theta(y_i \mid \mathbf{y}^{(t-1)}, q)$ and reveals a subset according to a *reveal order* π . The resulting states form a *denoising trajectory* $\tau^\pi = (\mathbf{y}^{(0)}, \dots, \mathbf{y}^{(T)})$. Standard inference uses a confidence-based order; LLaDA-8B (Nie et al., 2025) uses $T=128$ steps, Dream-7B (Ye et al., 2025) uses $T=64$.

Once revealed, a token is *committed*: all subsequent steps condition on it irreversibly. A factually incorrect early commitment acts as a *contextual attractor*, biasing neighbors toward internally consistent but wrong predictions via bidirectional attention (Chang et al., 2025; Qian et al., 2026). We call this *premature commitment hallucination*. Unlike autoregressive hallucination, confined to the generation frontier, a single early error propagates through all remaining denoising steps.

The key insight is that such hallucinations are *order-dependent*: a different reveal order $\pi' \neq \pi$ may commit a different token first, potentially yielding a correct output—so running N parallel chains with independently randomized reveal orders exposes this latent uncertainty without external supervision.

Definition 1 (Cross-Chain Entropy). Let $y_i^{(T,n)}$ denote the final token at position i from chain n . The empirical distribution is $\hat{p}_i(v) = \frac{1}{N} \sum_{n=1}^N \mathbb{1}[y_i^{(T,n)} = v]$. The **cross-chain entropy** at position i is:

$$H_{\times,i} = - \sum_{v \in \mathcal{V}} \hat{p}_i(v) \log \hat{p}_i(v) \quad (1)$$

Positions with $H_{\times,i}=0$ are *commitment-stable*: every chain agrees regardless of reveal order. Positions with high $H_{\times,i}$ are *commitment-uncertain*: the output depends on the arbitrary ordering rather than on factual knowledge. This signal is *model-native*, from the DLM’s denoising dynamics without a trained classifier or labeled data, and *structurally specific to DLMs*; autoregressive generation needs full resamples, while DLMs use parallel chains to isolate effects. With token vocabulary $|\mathcal{V}|$ over 32,000, the mean distinct tokens per position across $N=8$ chains is 2.7 (median 1, max 8), ensuring reliable entropy estimation. Cross-chain entropy, as a standalone detection signal, matches trained trajectory classifiers while requiring no supervision, capturing factual uncertainty directly (§5.1).

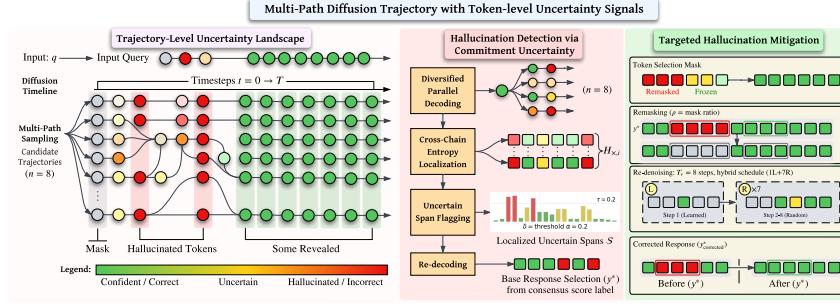


Figure 2: OSCAR overview. Phase 1: N parallel denoising chains with randomized reveal orders generate diverse trajectories. Cross-chain entropy $H_{\times}$ identifies high-uncertainty positions (red). Phase 2: targeted remasking, conditioned on retrieved evidence, corrects hallucinated spans at $1.3\times$ overhead.

3.2 Localization Formulation and CDH Metric

Definition 2 (Commitment Uncertainty Localization). Given query q and N trajectories $\{\tau^n\}_{n=1}^N$, identify:

$$\mathcal{U} = \{i : H_{\times,i} > Q_{1-\alpha}(H_{\times,1:L})\} \quad (2)$$

where $Q_{1-\alpha}$ is the $(1-\alpha)$ -quantile of the entropy distribution (default localization aggressiveness threshold $(\alpha)=0.2$, flagging the top 20% most uncertain positions).

This formulation is **training-free** ($Q_{1-\alpha}$ is computed from the current sample, no labeled data), **model-native** (the signal originates entirely from the DLM’s denoising process), and **pre-correction** (uncertainty is measured before any modification, avoiding post-hoc circularity).

To evaluate localization quality at the token level, we introduce the *cross-chain divergence-at-hallucination rate*:

$$\text{CDH}(k) = \frac{|\{i \in \mathcal{U}_k : i \text{ is hallucinated}\}|}{|\{i : i \text{ is hallucinated}\}|} \quad (3)$$

where $\mathcal{U}_k$ contains positions with the top- $k\%$ highest $H_{\times,i}$. $\text{CDH}(k)$ measures the fraction of truly hallucinated positions captured by the top- $k\%$ most uncertain positions (random baseline: $k/100$). Because OSCAR’s correction acts on localized spans, CDH directly measures the signal driving mitigation performance. We show in §5.3 that cross-chain entropy concentrates hallucinated positions in the high-uncertainty tail far more effectively than random selection, single-chain confidence remasking, and the strongest trained alternative, confirming that unsupervised localization from the model’s own trajectories is sufficient for targeted correction.

3.3 The OSCAR Pipeline

OSCAR implements a general program for commitment uncertainty localization in DLMs: (1) diversify reveal orders to expose latent factual uncertainty; (2) compute a position-level disagreement statistic from trajectories; (3) threshold to localize uncertain spans; (4) correct via targeted re-denoising conditioned on context or evidence. Alternative disagreement statistics (token-level KL divergence, embedding-space distance, optimal-transport-based discrepancies that compare distributions across multiple decoding paths (Mehta et al., 2023)) and stronger correction stages are possible; we present the choices yielding the strongest empirical results, and operationalize the program in three stages (Figure 2).

Stage 1: Diversified parallel decoding. OSCAR samples N reveal orders uniformly at random and runs N chains in parallel, sharing model weights θ . This use of batched parallel stochastic evaluations is conceptually related to scalable Monte Carlo inference, where

computational reparameterizations have been used to make large numbers of samples practical in Bayesian deep neural networks (Nazarovs et al., 2021). Standard confidence-based decoding always commits the most confident predictions first, suppressing observable uncertainty via contextual attractors (§3.1). Randomized orders break this coupling, ensuring no position receives systematic priority. All chains are batched into a single forward pass per step, yielding $\sim 1.2\times$ wall-clock overhead.

Stage 2: Localization and span aggregation. OSCAR computes $H_{\times,i}$ (Eq. 1) at each position and applies the percentile threshold (Eq. 2) to obtain the uncertain set $\mathcal{U}$. Consecutive positions are grouped into spans, each extended by $w=2$ tokens on either side; spans shorter than $\ell_{\min}=3$ tokens are discarded (ablated in Appendix B.2). The base response $\mathbf{y}^*$ is the chain whose tokens best match the cross-chain consensus:

$$n^* = \arg \max_n \sum_{i=1}^L \hat{p}_i(y_i^{(T,n)}) \quad (4)$$

Stage 3: Targeted remasking and correction. For each span $s = (a, \dots, b) \in \mathcal{S}$, OSCARremasks positions a through b in $\mathbf{y}^*$ and runs T_r denoising steps with all surrounding tokens frozen:

$$\hat{\mathbf{y}}_s = \arg \max_{\mathbf{y}_{a:b}} p_\theta(\mathbf{y}_{a:b} \mid \mathbf{y}_{\setminus s}^*, [\mathbf{e}_s; q]) \quad \text{via } T_r \text{ denoising steps} \quad (5)$$

where $\mathbf{e}_s$ is optional retrieved evidence. This exploits the DLM’s native ability to fill masked regions given surrounding context, an affordance unavailable in autoregressive models.

The first refinement step uses the learned confidence-based order, anchoring the correction to prevent a new attractor; the remaining $T_r - 1$ steps use random orders so the span settles without reproducing the original error. When a retrieval corpus is available, the uncertain span text (not the full input) serves as the query, focusing evidence on the claim under correction; we use Contriever (Chen et al., 2026) over a 2021 Wikipedia snapshot (top-1 passage, 256 tokens).

Overhead and hyperparameters. The correction stage remasks $\sim 18.7\%$ of tokens on average. With $T_r=8$ steps and batched inference, correction adds $\sim 0.1\times$; combined with diversification ($\sim 1.2\times$), total overhead is $\sim 1.3\times$. Defaults: $N=8$ chains, $\alpha=0.2$, $T_r=8$, $w=2$, $\ell_{\min}=3$, 1-Learned + (T_r-1) -Random schedule. Each is validated by ablations in later sections.

3.4 Crystallization: When Do Hallucinations Form?

The cross-chain entropy signal is not only diagnostic of *where* hallucinations occur but also *when* they form during denoising. We define the *entropy gap* at step t as the difference in mean cross-chain entropy between hallucinated and grounded positions:

$$\Delta H(t) = \mathbb{E}[H_{\times} \mid \text{hallucinated}, t] - \mathbb{E}[H_{\times} \mid \text{grounded}, t] \quad (6)$$

We find $\Delta H(t)$ ’s temporal profile is task-dependent. In factual QA the gap peaks early and persists—an early incorrect commitment dominates via bidirectional attention—while in summarization it grows mid-denoising, as compositional errors across spans require more steps to surface as cross-chain disagreement. Token-level ground truth uses RAGTruth’s span annotations; for TriviaQA, we align the final gold answer against each intermediate state via exact match. This pattern supports OSCAR: diverse reveal orders are most informative when commitment-order effects are strongest, and it suggests early-exit detection to reduce overhead (§6.4).

4 Experimental Setup

Models and datasets. We evaluate on LLaDA-8B-Instruct (Nie et al., 2025) ($T=128$ steps) and Dream-7B-Instruct (Ye et al., 2025) ($T=64$). Benchmarks: TriviaQA (Joshi et al., 2017)

Family	Method	LLaDA-8B	Dream-7B
Output	Perplexity	55.1	61.5
Output	LN-Entropy	60.8	58.0
Output	Semantic Entropy	65.4	62.8
Output	Lexical Similarity	54.9	66.5
Latent	EigenScore	63.1	69.8
Latent	CCS	68.6	54.1
Latent	TSV	70.4	65.2
Sampling	SelfCheckGPT-DLM	57.7	63.5
Trajectory	TraceDet [†]	72.0	80.8
Trajectory	DynHD [†]	84.2	84.3
Ours	OSCAR (detection-only, training-free)	<u>82.7</u>	<u>82.0</u>

[†] = requires a classifier trained on hallucination labels.

Table 1: **Unified-protocol detection AUROC (%)**. All methods scored under one method-blind GPT-4o judge prompt, averaged over TriviaQA, HotpotQA, and CommonsenseQA at both step counts. OSCAR here is the raw cross-chain entropy signal with *no correction applied*. Best **bolded**, second underlined; significance tests in §5.1, per-dataset numbers in Appendix C.1.

(500), HotpotQA (Yang et al., 2018) (500), CommonsenseQA (Talmor et al., 2019) (500; negative control—multiple-choice format yields $H_{\times} \approx 0$), and RAGTruth (Niu et al., 2024) (500; token-level span annotations).

Baselines. Eleven baselines across four families: output-based (Perplexity, LN-Entropy, Semantic Entropy (Kuhn et al., 2023; Quevedo et al., 2025), Lexical Similarity (Manakul et al., 2023)), latent-based (EigenScore (Shoushtari et al., 2025), CCS (Burns et al., 2022), TSV (Chang et al., 2025)), sampling-based (SelfCheckGPT-DLM: cross-generation disagreement over $N=8$ independently sampled diffusion generations without reveal-order diversification), and trajectory-based (TraceDet (Chang et al., 2025), DynHD (Qian et al., 2026)), plus token-level majority vote across $N=8$ chains. For mitigation and localization we additionally implement **ReMDM-conf**, a post-hoc confidence-based remasking baseline capturing the core mechanism of the inference-time remasking literature (Ghazvininejad et al., 2019; Lezama et al., 2022; Wang et al., 2025; Huang et al., 2026): tokens are selected via single-chain per-token confidence, then re-denoised under conditions matched to OSCAR exactly (18.7% budget, $w=2$, $\ell_{\min}=3$, $T_r=8$).

Unified evaluation protocol. All methods, including the trained detectors TraceDet and DynHD are scored under a single method-blind GPT-4o (Achiam et al., 2023) LLM-as-Judge prompt (Appendix E), with exact match (EM) as a transparent lexical lower bound; judge reliability is validated via label binarization, a second judge, and a human-labeled subset (§5.1, Appendix D.1). We report AUROC, F1 before/after correction, CDH(k), and RAGTruth span reduction, as mean $\pm$ std over three seeds. Defaults: $N=8$, $\alpha=0.2$, $T_r=8$; batched on $4 \times \text{H200 80GB}$; retrieval via Contriever (Chen et al., 2026) over Wikipedia.

5 Discussion

5.1 Hallucination Detection: Matching Trained Detectors Without Supervision

Table 1 presents the headline detection comparison under the unified judge protocol. As a pure detector, cross-chain entropy only, no correction applied OSCAR attains **82.7** AUROC on LLaDA-8B and **82.0** on Dream-7B: statistically tied with DynHD (the strongest trained detector), above TraceDet on LLaDA-8B by +10.7, and above every output-, latent-, and sampling-based baseline by 12–28 points, with *no hallucination-specific training, labeled data, or learned reference model*. We make no claim of surpassing trained detectors at detection; the contribution is *matched detection without supervision, plus a correction stage that no detector baseline provides* (§5.2). EM evaluation, a transparent lexical lower bound, yields 76.4/81.8;

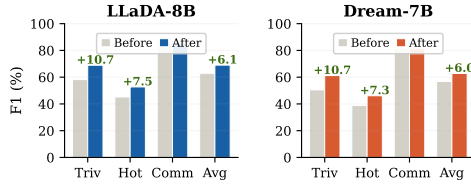


Figure 3: F1 before (gray) and after (color) OSCAR correction. Green annotations show $\Delta F1$. CommonsenseQA shows no change ($H_{\times} \approx 0$), validating selectivity. Mean $\pm$ std over 3 seeds; full metrics in Appendix B.4.

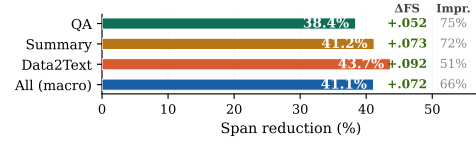


Figure 4: Span-level correction on RAGTruth (LLaDA-8B). Bars show hallucinated span reduction; ΔFS and % Improved annotated at right. Full per-subset metrics in Appendix A.

the trained detectors’ EM-label advantage disappears under semantic evaluation. After correction, judge-reabeled *full-pipeline* AUROC rises to 86.5/85.7 (Appendix C.1, Table 15); we report this separately from Table 1 since it reflects post-correction outputs, and a detector-to-pipeline comparison would conflate the two contributions.

Judge reliability. Conclusions are robust to the evaluation instrument: a second, architecturally distinct judge (Claude Opus 4.1) preserves method ordering (judge $\leftrightarrow$ judge $\kappa=0.74$), and on a human-labeled subset ($N=200$) GPT-4o tracks human semantic judgment more closely than exact match does ($\kappa=0.76$ vs. 0.61), with OSCAR $\leftrightarrow$ DynHD remaining a statistical tie under human labels ($\Delta=+0.4$, n.s.)—so OSCAR’s standing is not an artifact of the judge (full protocol and binarization checks in Appendix D.1).

5.2 Generation Quality and Span Correction

Figure 3 shows F1 before and after correction: OSCAR improves QA macro-average F1 by +6.1 pp on LLaDA-8B and +8.3 pp across both models, largest on TriviaQA (+10.7), with Dream-7B showing similar gains. On CommonsenseQA, $\Delta F1 = 0.0$: chains agree on single-token answers ($H_{\times} \approx 0$) so OSCAR doesn’t intervene, and its strong sample-level AUROC there shows these confident outputs are factual; the signals are complementary, not contradictory. Figure 4 shows correction extends beyond short-form QA: hallucinated spans drop 38–44% across RAGTruth subsets with positive ΔFS , and uncertainty-targeted retrieval adds +3.4 F1 (Appendix A.1). Three analyses confirm the corrections are genuine, not evaluation artifacts: (i) held-out perplexity is statistically unchanged after correction ($p \geq 0.18$) with fluency grades moving ≤ 1 point—coherent rewrites, not disfluent swaps that game the judge (Appendix B.5); (ii) on an ambiguous/multi-answer stress set ($N=150$), OSCAR over-writes a valid alternative in only 5.8% of cases, no higher than its 6.2% general degradation rate (Appendix B.7); (iii) among residual failures, failed remasking accounts for only 6–18%, with knowledge gaps and missed localization dominating the localize-and-remask mechanism itself is sound (Appendix B.8).

5.3 Localization Quality: CDH Analysis

Appendix A.2 reports CDH(k) curves. At $k=20\%$ (i.e., examining only the top 20% most uncertain positions), OSCAR’s cross-chain entropy captures 67.3% of all hallucinated positions on LLaDA-8B—more than $3\times$ the random baseline (20%), $1.4\times$ the best trained detector (TraceDet, 47.8%), and $2.4\times$ single-chain confidence remasking (ReMDM-conf, 27.6%). Single-chain confidence localizing barely above random empirically demonstrates the contextual-attractor effect of §3.1: an early wrong commitment biases neighbors to be *confidently* wrong, so the error cannot be surfaced from within one trajectory, the gap reveal-order diversification closes. On mitigation under matched conditions, ReMDM-conf likewise trails OSCAR on every metric (Triv+Hot F1 55.5 vs. 60.8; span reduction 23.4% vs. 41.1%; ΔFS +0.034 vs. +0.072; Table 5, Appendix A.2): cross-chain entropy is a distinct signal class, not reducible to confidence-based remasking.

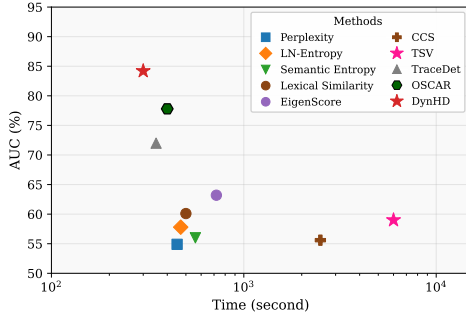


Figure 5: **Pareto frontier**. AUC = LLaDA-8B average. $\star$ = Pareto-optimal. Wall-clock on $4\times H200$.

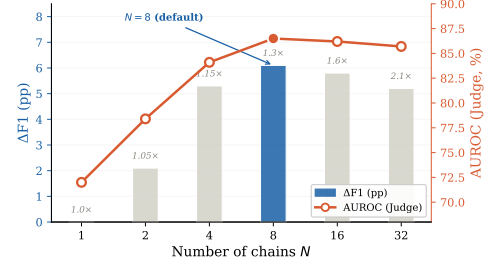


Figure 6: Effect of chains N on LLaDA-8B (QA macro-avg).

5.4 Pareto Analysis: Performance vs. Efficiency

Figure 5 (left) presents the Pareto frontier. The full OSCAR pipeline achieves the highest post-correction AUROC at $1.3\times$ wall-clock overhead, establishing a Pareto-optimal point for the combined detection-plus-mitigation setting; as a detector alone, OSCAR is tied with DynHD (Table 1) while requiring no training. The key insight: even with $N=8$ parallel chains, batched inference keeps the overhead manageable, while the detection + correction synergy produces gains that neither component achieves alone.

6 Ablation and Analysis

We decompose OSCAR’s gains to identify which components are load-bearing, how sensitive the method is to its control parameters, and why cross-chain entropy is an effective uncertainty signal.

6.1 Localization and Correction are Complementary

Table 2 shows that localization alone does not affect F1 performance, while random span corrections slightly reduce results. The complete OSCAR pipeline integrates localization with targeted correction, achieving a 6.1-point F1 improvement and 86.5 post-correction AUROC, outperforming untargeted correction (+2.5 F1) and token-level majority voting (+2.2 F1). The default demasking schedule (1L+7R) likewise outperforms every alternative ordering scheme across all metrics (Table 14, Appendix B.10): localization and targeted correction are synergistic.

6.2 Compute-Matched Isolation of the Reveal-Order Mechanism

Does OSCAR’s signal derive from reveal-order trajectory construction, or simply from more forward passes? Holding total compute fixed (64 forward passes, LLaDA-8B) and varying

Configuration	F1 (%)	AUROC	$\Delta F1$
Unguided decoding	62.9	—	—
Best-of-N selection (no correction)	62.9	82.7	+0.0
Correction only (random spans)	62.8	—	−0.1
Majority vote — token	65.1	—	+2.2
Localization + untargeted correction	65.4	83.1	+2.5
OSCAR (loc. + targeted)	69.0	86.5[†]	+6.1

Table 2: **Complementarity ablation** (LLaDA-8B, QA macro-avg). [†]Full pipeline under LLM-as-Judge (post-correction); detection-only AUROC = 82.7 (raw entropy signal), EM = 76.4.

only chain construction and disagreement statistic (Table 17, Appendix C.2), reveal-order diversification alone accounts for +19.1 AUROC / +3.8 F1 (rows 1→2) the dominant factor and switching the statistic from confidence to cross-chain entropy adds a further +5.9 / +2.1 (rows 2→3); independent sampling without reveal-order is weakest (57.7, matching SelfCheckGPT-DLM, Appendix C.2). This isolates reveal-order trajectory construction—not forward-pass budget—as the signal’s source.

6.3 Sensitivity Analysis

Figure 6 shows the effect of chains N on LLaDA-8B (QA macro-avg): F1 gain and AUROC rise to a peak at $N = 8$ and plateau at $N = 16\text{--}32$, making $N = 8$ the quality/cost knee. For demasking order, the hybrid schedule (one learned step, then $(T_r - 1)$ random) beats all-learned (61.8 F1, below the 62.9 baseline), all-random (64.6), and entropy-ordered (63.2): the learned step anchors the most confident token while random steps add diversity (Table 14). Corrections show 91–97% precision (Appendix B.1), and α is optimal across $[0.15, 0.25]$, degrading at extremes (Appendix B.2).

6.4 Why Does Cross-Chain Entropy Work?

Cross-chain entropy exploits an architectural asymmetry: in AR decoding, uncertainty at position i is hidden at $i + 1$ because only the argmax token propagates, and recovering it requires resamples that mix sampling variance with factual uncertainty; DLMs maintain full distributions, so different reveal orders sample different paths and cross-chain entropy measures the width of this path distribution per position, exactly why an unsupervised signal from $N = 8$ chains matches classifiers trained on trajectory features (Table 1). The hand-audit of flagged disagreements (Appendix B.6) confirms the signal is predominantly factual (71% factual errors, 21% paraphrase, 8% formatting), with the non-factual minority left intact by the consensus-anchored correction. Confident-but-wrong cases, where all chains agree on an incorrect answer ($H_\times = 0$), represent 9–20% of hallucinated positions (Appendix B.3).

7 Conclusion

We introduced commitment uncertainty localization, formalized through the CDHmetric, and OSCAR, a training-free framework that runs N parallel chains with randomized reveal orders, localizes unstable positions via cross-chain entropy, and corrects them via targeted remasking. On LLaDA-8B and Dream-7B, OSCAR improves F1 by $+8.3 \pm 0.4$, reduces hallucinated spans by 41.1%, matches trained detectors under unified judge protocol, and reaches 86.5 post-correction AUROC at $1.3\times$ overhead. Structural asymmetry lets DLMs maintain full distributions throughout denoising, enabling inference-time localization and correction absent in AR generation.

Future work. The crystallization analysis motivates early-exit detection and lower-cost correction schedules; extending the localization to continuous or hybrid DLMs would test the signal’s generality; and adaptive policies like task-dependent α , span selection, adaptive chain counts (Appendix B.9), may further improve the accuracy–efficiency trade-off; integration with preference optimization, retrieval, or tool use could broaden OSCAR’s role further. We release the trajectory divergence toolkit to support such work.

Limitations

Six limitations. (1) OSCAR runs N parallel chains, raising VRAM $\sim 1.67\times$ at $N=8$; $N=4$ retains $\sim 80\%$ gain when VRAM-constrained. (2) If the model lacks knowledge, remasking cannot help: chains hallucinate consistently (9–20% of positions; Appendix B.3). (3) LLM-as-Judge evaluation is biased despite mitigation; gains remain judge-reported only. (4) Generalization beyond LLaDA-8B and Dream-7B is unguaranteed. (5) span hyperparameters are empirical. (6) adaptive- N is proposed, not evaluated (Appendix B.9).

Ethics Statement

This work proposes training-free inference-time techniques for reducing hallucination in diffusion language models. We do not introduce new training data, fine-tuned models, or large-scale data collection. The models evaluated (LLaDA-8B, Dream-7B) are publicly available research artifacts; evaluations use public benchmarks (TriviaQA, HotpotQA, CommonsenseQA, RAGTruth). LLaDA-8B and Dream-7B are the only publicly available DLMs at scale as of early 2026; our evaluation therefore covers the full available ecosystem. The human-agreement study (Appendix D.1) used two NLP-trained annotators labeling model outputs for factual correctness; no personal or sensitive data was involved. Technical limitations, including computational overhead, failure modes under complete knowledge absence, evaluation biases, and hyperparameter sensitivity are detailed in §7. The LLM-as-Judge evaluation (GPT-4o, with Claude Opus 4.1 as a robustness check) introduces potential biases; a human-labeled subset partially validates the judge, but no human evaluation of user-perceived helpfulness is included. GPT-4o assisted in polishing this manuscript; associated costs and biases are acknowledged.

Acknowledgments

Prof. Lokhande acknowledges support from University at Buffalo startup funds, an Adobe Research Gift, an NVIDIA Academic Grant, and the National Center for Advancing Translational Sciences of the NIH (award UM1TR005296 to the University at Buffalo).

References

- OpenAI Josh Achiam, Steven Adler, S. Agarwal, L. Ahmad, Ilge Akkaya, Florencia Leoni Aleman, D. Almeida, Janko Altenschmidt, S. Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv*, 2023.
- Amos Azaria and Tom M. Mitchell. The internal state of an llm knows when its lying. *Conference on Empirical Methods in Natural Language Processing*, abs/2304.13734:967–976, 2023. doi: 10.18653/v1/2023.findings-emnlp.68. URL <https://api.semanticscholar.org/CorpusID:258352729>.
- Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks. *Journal of machine learning research*, 21(1), 2020. ISSN 1532-4435.
- Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *International Conference on Learning Representations*, abs/2212.03827, 2022. doi: 10.48550/arXiv.2212.03827. URL <https://api.semanticscholar.org/CorpusID:254366253>.
- Shenxu Chang, Junchi Yu, Weixing Wang, Yongqiang Chen, Jialin Yu, Philip H. S. Torr, and Jindong Gu. Tracedet: Hallucination detection from the decoding trace of diffusion large language models. *arXiv.org*, abs/2510.01274, 2025. doi: 10.48550/arXiv.2510.01274. URL <https://api.semanticscholar.org/CorpusID:281724918>.
- Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llm’s internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, volume 2024, pp. 3056–3076, 2024. doi: 10.48550/arXiv.2402.03744.
- Haitian Chen, Qingyao Ai, Xiao Wang, Yiqun Liu, Fen Lin, and Qin Liu. Unsupervised dense retrieval with counterfactual contrastive learning. *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, pp. 47–57, 2026. doi: 10.1145/3773966.3778015.
- Sotirios Panagiotis Chytas, Vishnu Lokhande, and Vikas Singh. Pooling image datasets with multiple covariate shift and imbalance. In *International Conference on Learning Representations*, volume 2024, pp. 57282–57304, 2024. doi: 10.48550/arXiv.2403.02598.

- Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In *Conference on Empirical Methods in Natural Language Processing*, pp. 6111–6120. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1633.
- Zemin Huang, Yuhang Wang, Zhiyang Chen, and Guo-Jun Qi. Don’t settle too early: Self-reflective remasking for diffusion language models. In *arXiv.org*, volume 2026, pp. 56998–57026, 2026. doi: 10.48550/arXiv.2509.23653.
- Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabrys, Ravid Shwartz-Ziv, and Tomasz Kajdanowicz. The illusion of progress: Re-evaluating hallucination detection in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34716–34733. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1761. URL <https://api.semanticscholar.org/CorpusID:280635762>.
- Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Annual Meeting of the Association for Computational Linguistics*, abs/1705.03551:1601–1611, 2017. doi: 10.18653/v1/P17-1147. URL <https://api.semanticscholar.org/CorpusID:26501419>.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *International Conference on Learning Representations*, abs/2302.09664, 2023. doi: 10.48550/arXiv.2302.09664. URL <https://api.semanticscholar.org/CorpusID:257039062>.
- Patrick Lewis, Ethan Perez, Aleksandara Piktus, F. Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, M. Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Neural Information Processing Systems*, 33:9459–9474, 2020.
- José Lezama, Huiwen Chang, Lu Jiang, and Irfan Essa. Improved masked image generation with token-critic. In *Lecture Notes in Computer Science*, pp. 70–86. Springer, Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-20050-2_5.
- Tianyi Li, Mingda Chen, Bowei Guo, and Zhiqiang Shen. A survey on diffusion language models. *arXiv.org*, 2025. doi: 10.48550/arXiv.2508.10875.
- Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Trans. Mach. Learn. Res.*, 2023. doi: 10.48550/arXiv.2305.19187.
- Rui Lu, Runzhe Wang, Kaifeng Lyu, Xitai Jiang, Gao Huang, and Mengdi Wang. Towards understanding text hallucination of diffusion models via local generation bias. In *International Conference on Learning Representations*, volume 2025, pp. 24192–24219, 2025. doi: 10.48550/arXiv.2503.03595.
- Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *International Conference on Learning Representations*, 2021. doi: 10.17863/CAM.63497.
- Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.557.
- Ronak Mehta, Jeffery Kline, {Vishnu Suresh} Lokhande, Glenn Fung, and Vikas Singh. Efficient discrete multi-marginal optimal transport regularization. 2023. Publisher Copyright: © 2023 11th International Conference on Learning Representations, ICLR 2023. All rights reserved.; 11th International Conference on Learning Representations, ICLR 2023 ; Conference date: 01-05-2023 Through 05-05-2023.

- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.741.
- Jurijs Nazarovs, Ronak R Mehta, Vishnu Suresh Lokhande, and Vikas Singh. Graph reparameterizations for enabling 1000+ monte carlo iterations in bayesian deep neural networks. In *Conference on Uncertainty in Artificial Intelligence*, pp. 118–128. PMLR, 2021.
- Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *Advances in Neural Information Processing Systems* 38, 38:56354–56392, 2025. doi: 10.52202/085713-1689.
- Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.585.
- Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. Steer llm latents for hallucination detection. *International Conference on Machine Learning*, 2025.
- Ya Ru Qian, Yue Tan, Yixin Liu, Wang Yu, and Shirui Pan. Dynhd: Hallucination detection for diffusion large language models via denoising dynamics deviation learning. *arXiv.org*, abs/2603.16459, 2026. doi: 10.48550/arXiv.2603.16459. URL <https://api.semanticscholar.org/CorpusID:286583763>.
- Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomás Cerný. Detecting hallucinations in large language model generation: A token probability approach. *Communications in Computer and Information Science*, abs/2405.19648:154–173, 2025. doi: 10.1007/978-3-031-86623-4_13. URL <https://api.semanticscholar.org/CorpusID:270123170>.
- Jie Jessie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. Out-of-distribution detection and selective generation for conditional language models. *International Conference on Learning Representations*, abs/2209.15558, 2022. doi: 10.48550/arXiv.2209.15558. URL <https://api.semanticscholar.org/CorpusID:252668614>.
- Shirin Shoushtari, Yi Wang, Xiao Shi, M. Salman Asif, and Ulugbek S. Kamilov. Eigenscore: Ood detection using covariance in diffusion models. *arXiv.org*, abs/2510.07206, 2025. doi: 10.48550/arXiv.2510.07206. URL <https://api.semanticscholar.org/CorpusID:281891561>.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *North American Chapter of the Association for Computational Linguistics*, 2019. doi: 10.18653/v1/N19-1421.
- SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv.org*, 2024. doi: 10.48550/arXiv.2401.01313.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Neural Information Processing Systems*, 30, 2017. doi: 10.65215/ctdc8e75.
- Guanghan Wang, Yair Schiff, Subham Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. *Advances in Neural Information Processing Systems* 38, 38:163488–163545, 2025. doi: 10.52202/085713-4927.

- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. 2023. URL <https://openreview.net/forum?id=1PL1NIMMrw>.
- 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 *Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL <https://api.semanticscholar.org/CorpusID:52822214>.
- Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv.org*, abs/2508.15487, 2025. doi: 10.48550/arXiv.2508.15487. URL <https://api.semanticscholar.org/CorpusID:280700361>.

Appendix

Table of Contents

A	RAGTruth: Span-Level Results	16
	A.1 Retrieval Augmentation	16
	A.2 CDH Localization Curves	16
B	Ablation and Robustness Studies	17
	B.1 Correction Precision	17
	B.2 Sensitivity: α , T_r , Span, N -Paths	17
	B.3 Confident-but-Wrong Analysis	18
	B.4 Extended Generation Metrics	18
	B.5 Fluency of Corrected Outputs	18
	B.6 Disagreement-Type Decomposition	18
	B.7 Ambiguous / Multi-Answer Stress Test	19
	B.8 Failure-Mode Breakdown	19
	B.9 Adaptive- N Chain Allocation	20
	B.10 Demasking Order for Correction	20
C	Baseline Comparisons	20
	C.1 Per-Dataset Breakdown	20
	C.2 SelfCheckGPT-DLM Comparison	20
D	Statistical Validation and Judge Reliability	20
	D.1 Bootstrap Confidence Intervals	20
	D.2 Judge Reliability and Human Agreement	20
E	Implementation Details	23
	E.1 LLM-as-Judge Prompt	23

A RAGTruth: Span-Level Results

Table 3 provides the full per-subset breakdown behind Figure 4. Span reduction is highest on Data2Text (43.7%), where structured facts produce clear-cut uncertain positions; the highest fraction of improved examples is on QA (75.0%), where single-span corrections have the largest relative impact. Positive Δ FS across all subsets confirms that corrections improve factual precision rather than simply deleting content.

Subset	N	Hall.%	AUROC	Δ FS	Span Red. %
QA	788	18.4	0.612	+0.052	38.4
Summary	500	24.2	0.641	+0.073	41.2
Data2Text	500	66.8	0.663	+0.092	43.7
All	1788	33.1	0.639	+0.072	41.1

Table 3: Span-level correction on RAGTruth (LLaDA-8B). Δ FS: FactScore change (positive = more precise). Span Red.: reduction in hallucinated span mass.

A.1 Retrieval Augmentation

Table 4 isolates the effect of retrieval strategy. OSCAR without retrieval (60.8) already exceeds naive RAG (53.5) by 7.3 points, the primary gain is from targeted remasking, not retrieved evidence. Adding targeted span-level retrieval yields a further +3.4 over OSCAR alone, because evidence is fetched for each uncertain claim rather than the full query.

Configuration	F1 (%)	Δ vs. Unguided
Unguided decoding	51.7	—
Naive RAG	53.5	+1.8
OSCAR (no retrieval)	60.8	+9.1
OSCAR + Naive RAG	62.4	+10.7
OSCAR + Targeted RAG	64.2	+12.5

Table 4: Retrieval comparison (LLaDA-8B, TriviaQA + HotpotQA avg F1).

A.2 CDH Localization Curves

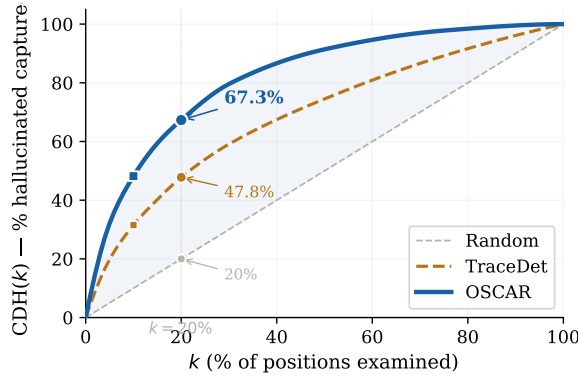


Figure 7: $CDH(k)$ on RAGTruth. At $k=20\%$, OSCAR captures 67.3% of hallucinated positions vs. 47.8% (TraceDet), 27.6% (ReMDM-conf), and 20% (random).

Figure 7 shows that hallucinated positions concentrate in the high-entropy tail: examining only the top 20% of positions by cross-chain entropy captures two-thirds of all true errors, making targeted remasking practical without modifying the remaining 80% of tokens. At $k=10\%$, OSCAR already captures 48.2% (vs. 31.5% TraceDet, 10% random), confirming that

the unsupervised signal localizes hallucinations more effectively than both trained and single-chain-confidence alternatives.

Method	CDH(20%)	Triv+Hot F1	Span Red.%	Δ FS
Random baseline	20.0	—	—	—
ReMDM-conf (single-chain confidence)	27.6	55.5	23.4	+0.034
TraceDet (trained)	47.8	—	—	—
OSCAR (cross-chain entropy)	67.3	60.8	41.1	+0.072

Table 5: **Localization and mitigation vs. confidence-based remasking** (LLaDA-8B). ReMDM-conf matches OSCAR’s remasking budget (18.7%), span aggregation ($w=2, \ell_{\min}=3$), and refinement steps ($T_r=8$) exactly; only the localization criterion differs. Mitigation numbers exclude retrieval for both methods.

B Ablation and Robustness Studies

B.1 Correction Precision

Among examples where OSCAR applies a correction, 91.9% are improved on TriviaQA (57 corrected, 5 broken out of 500) and 97.1% on HotpotQA (34 corrected, 1 broken). CommonsenseQA triggers no intervention.

B.2 Sensitivity: α , T_r , Span, N -Paths

α	% Remasked	F1 (%)	AUROC
0.05	5%	64.1	83.8
0.10	10%	66.5	85.1
0.15	15%	68.2	86.0
0.20	19%	69.0	86.5
0.25	25%	68.1	86.3
0.30	30%	66.8	85.9
0.40	40%	63.7	84.2

Table 6: Threshold α sensitivity (LLaDA-8B, QA macro-avg).

Setting	F1 (%) / Δ FS	Cost
<i>Refinement steps T_r</i>		
$T_r=2$	65.4 / +0.041	0.03×
$T_r=4$	67.8 / +0.058	0.06×
$T_r=8$	69.0 / +0.072	0.10×
$T_r=16$	69.2 / +0.074	0.19×
<i>Span aggregation ($w, \ell_{\min}$)</i>		
$w=0, \ell=1$	65.8	28.4% red.
$w=1, \ell=3$	67.4	35.2%
$w=2, \ell=3$	69.0	41.1%
$w=3, \ell=3$	68.3	45.8%
$w=2, \ell=5$	67.9	37.6%
<i>N-paths FactScore (RAGTruth)</i>		
$N=1$	Δ FS = −0.030	—
$N=4$	+0.073	—
$N=8$	+0.148	—
$N=16$	+0.048	—

Table 7: Refinement steps T_r , span aggregation, and N -paths FactScore (LLaDA-8B). Defaults bolded.

α plateaus across $[0.15, 0.25]$; F1 and Δ FS plateau at $T_r=8$; span defaults ($w=2, \ell_{\min}=3$) achieve peak F1; FactScore peaks at $N=8$.

B.3 Confident-but-Wrong Analysis

Dataset	Total Hall.	CBW	Detectable	CBW Rate
TriviaQA	245	38	207	15.5%
HotpotQA	315	63	252	20.0%
RAGTruth	592	53	539	9.0%

Table 8: CBW hallucinations (LLaDA-8B). $CBW = H_{\times,i}=0$ but factually wrong.

80–91% of hallucinated positions expose non-zero cross-chain entropy and are detectable; the remaining CBW cases represent knowledge gaps requiring retrieval rather than remasking. The higher HotpotQA rate (20%) reflects two-hop structure: when both bridging facts are missing, all reveal orders converge; the lower RAGTruth rate (9%) reflects more reveal-order-dependent compositional errors.

B.4 Extended Generation Metrics

Dataset	EM		F1		R-L		BLEU	
	Bef.	Aft.	Bef.	Aft.	Bef.	Aft.	Bef.	Aft.
<i>LLaDA-8B</i>								
TriviaQA	51.0	61.4	58.2	68.9±0.3	55.1	65.4	33.2	41.8
HotpotQA	37.0	43.6	45.2	52.7±0.4	42.8	50.1	24.6	30.1
CommQA	85.4	85.4	85.4	85.4±0.0	85.4	85.4	72.1	72.1
<i>Dream-7B</i>								
TriviaQA	42.0	53.8	50.5	61.2±0.5	48.3	58.7	—	—
HotpotQA	30.2	37.4	38.8	46.1±0.6	36.5	43.7	—	—
CommQA	81.2	81.2	81.2	81.2±0.0	81.2	81.2	—	—

Table 9: Full generation metrics before/after OSCAR correction (mean ± std over 3 seeds).

B.5 Fluency of Corrected Outputs

Targeted remasking rewrites a fixed-length window conditioned on context that was committed assuming the original content, raising the possibility of boundary discontinuities that a permissive factuality judge would not penalize. Two design choices target continuity: span extension ($w=2$) and the 1-Learned+ (T_r-1) -Random schedule, whose first learned step anchors the correction to its committed neighbors (Table 14). Table 10 reports per-example held-out perplexity and a fluency-only GPT-4o grade in which factual correctness is deliberately ignored, on the longest-form outputs in the current public DLM ecosystem (RAGTruth Summary/Data2Text, multi-sentence, avg. 147 tokens).

Subset	Held-out PPL orig→corr	Paired p	% fluent orig→corr
QA (Triv+Hot)	24.7 → 24.2	0.18	94 → 93
RAGTruth Summary	29.4 → 29.0	0.27	91 → 90

Table 10: Fluency of corrected outputs (LLaDA-8B). Perplexity is statistically unchanged and the fluency-only grade moves ≤ 1 point while factual scores rise: corrections are coherent rewrites, not disfluent token swaps that game the factuality judge.

B.6 Disagreement-Type Decomposition

Cross-chain disagreement could in principle arise from valid paraphrases, alternative answer forms, or unstable tokenization rather than factual uncertainty. We hand-audited $N=250$ flagged high-entropy token-level disagreements (150 from TriviaQA+HotpotQA, 100

from RAGTruth), each classified into one of three buckets by two independent annotators (inter-annotator $\kappa=0.73$).

Subset (N)	Factual error	Paraphrase / alt-form	Formatting / tokenization
TriviaQA + HotpotQA (150)	72%	21%	7%
RAGTruth (100)	69%	22%	9%
Overall (250)	71%	21%	8%

Table 11: Hand-audit of flagged high-entropy disagreements (LLaDA-8B).

The majority of flagged disagreement (71%) is genuinely factual. The minority lexical/format component (29%) is largely left intact by correction: the consensus-anchored base response (Eq. 4) retains the most-agreed chain, and correction precision is 91.9–97.1% (Appendix B.1), so corrections rarely overwrite a valid paraphrase. Cross-chain entropy measures commitment uncertainty, not factual error directly; but the dominant source of disagreement in factual QA is factual, and the non-factual component does not systematically trigger false corrections.

B.7 Ambiguous / Multi-Answer Stress Test

On ambiguous or multi-answer questions, high entropy may reflect legitimate answer diversity rather than hallucination. Three structural guards limit over-correction: (i) OSCAR does not intervene when chains agree (CommonsenseQA: $H_{\times} \approx 0$, zero interventions); (ii) the consensus-anchored base response (Eq. 4) retains the most-agreed valid answer; (iii) the 1-Learned anchor step preserves a confident valid answer during re-denoising. We measure over-correction directly on a hand-built ambiguous/alias set:

Subset	N	Valid alt. (before)	Over-written	Rate
TriviaQA aliases	84	61	3	4.9%
HotpotQA multi-hop	66	42	3	7.1%
Overall	150	103	6	5.8%

Table 12: Over-correction on ambiguous / multi-answer questions (LLaDA-8B). Rate = valid alternatives over-written $\div$ examples with a correct valid alternative before correction.

OSCAR over-writes a valid alternative in only 5.8% of cases, which is not systematically higher than the 6.2% general degradation rate (6 broken / 97 intervened across TriviaQA+HotpotQA, Appendix B.1): OSCAR does not have a specific over-correction problem on ambiguous questions relative to its general correction behavior.

B.8 Failure-Mode Breakdown

On the full 500-example splits, 61/67/55 examples remain judge-incorrect after the full pipeline on TriviaQA/HotpotQA/RAGTruth. (This differs from Appendix B.1’s 5/1 broken examples, which count only cases where OSCAR worsened a previously correct answer among intervened examples; here we attribute *all* post-pipeline failures.)

Cause	TriviaQA	HotpotQA	RAGTruth
Wrong / missed localization	41%	36%	33%
Failed remasking	10%	6%	18%
Knowledge gap / retrieval	49%	58%	49%

Table 13: Attribution of residual failures after the full pipeline (LLaDA-8B; example-level, $N=61/67/55$).

The correction mechanism itself is rarely at fault (failed remasking 6–18%); residual errors are dominated by knowledge gaps ($\rightarrow$ retrieval) and missed localization ($\rightarrow$ higher recall). The localize-and-remask mechanism itself is sound.

B.9 Adaptive- N Chain Allocation

Reviewer feedback suggested allocating chains adaptively: easy examples stop after fewer chains, uncertain examples receive more. Existing results support feasibility—Figure 6 shows a quality / compute knee at $N=8$ with meaningful signal from $N=4$, the crystallization analysis (§3.4) shows the entropy gap peaks early in factual QA, and CommonsenseQA already demonstrates the principle in the strongest sense (chains agree, zero corrections applied). We specify an explicit adaptive policy: generate $N=4$ chains, and escalate to $N=8$ only when the fraction of positions with $H_\times > 0$ exceeds a task-agnostic trigger. A full quantitative evaluation of this policy—stopping criterion, average chain count per task, and the F1/AUROC vs. fixed- N trade-off—is left to future work.

B.10 Demasking Order for Correction

Order	Triv. F1	CQ F1	HQA F1	RT R-L	AUROC	Avg F1
All learned	56.8	85.4	43.1	33.8	0.541	61.8
All random	61.2	85.4	47.3	37.4	0.608	64.6
Hybrid (50/50)	64.8	85.4	49.8	40.1	0.627	66.7
Entropy-ordered	59.4	85.4	44.7	35.2	0.591	63.2
OSCAR (1L+7R)	68.9	85.4	52.7	43.2	0.639	69.0

Table 14: **Demasking order ablation** (LLaDA-8B). OSCAR default = 1L+7R. The all-learned schedule falls below the 62.9 unguided baseline because it reproduces the original contextual attractor; the single learned anchor step plus $(T_r - 1)$ random steps combines stability with reveal-order diversity.

C Baseline Comparisons

C.1 Per-Dataset Breakdown

Table 15 reports the per-dataset breakdown. For output- and latent-based baselines, values are AUROC under EM labels re-evaluated on our DLM outputs under the same pipeline; OSCAR (EM) is the detection signal under EM labels; OSCAR (Judge) rows are the *full-pipeline*, *post-correction* scores under judge labels. The headline detector-to-detector comparison under the unified judge protocol is Table 1 in the main text.

Table 16 additionally reports the fair-baseline re-evaluation on TriviaQA and HotpotQA under identical conditions.

All baselines re-evaluated on our DLM outputs under the same pipeline.

C.2 SelfCheckGPT-DLM Comparison

OSCAR’s detection signal outperforms the best SelfCheckGPT variant by 18.7 AUROC points at matched N ; reveal-order diversification exposes a qualitatively different signal than independent resampling, consistent with the compute-matched ablation (Table 17).

D Statistical Validation and Judge Reliability

D.1 Judge Reliability and Human Agreement

Second judge. Re-scoring all methods with an architecturally distinct judge (Claude Opus 4.1) shifts method-specific AUROC by $-0.3/-0.9/-0.5$ with method ordering preserved; judge $\leftrightarrow$ judge agreement is $\kappa=0.74$.

Method		TriviaQA		HotpotQA		CSQA		Avg
		128	64	128	64	128	64	
LLaDA-8B-Instruct								
Output	Perplexity	50.4	47.6	49.3	51.2	65.6	65.0	54.9
	LN-Entropy	54.6	53.5	54.8	54.7	64.6	64.4	57.8
	Semantic Entropy	68.9	67.3	57.6	53.8	44.1	43.9	55.9
	Lexical Similarity	62.5	59.0	64.2	57.1	57.3	60.7	60.1
Latent	EigenScore	69.2	66.9	64.7	59.2	58.5	60.6	63.2
	CCS	57.1	54.2	57.6	55.8	50.5	58.5	55.6
	TSV	60.2	61.1	65.0	59.4	52.9	55.2	59.0
Traj.	TraceDet [†]	73.9	74.1	66.1	63.7	77.2	77.1	72.0
	DynHD [†]	86.7	86.1	84.2	85.3	81.6	81.3	84.2
	OSCAR (EM)	80.3	79.8	71.5	68.2	79.4	78.9	76.4
	OSCAR full pipeline (Judge)	89.7	88.8	86.7	87.5	83.4	82.8	86.5
Dream-7B-Instruct								
Out.	Semantic Entropy	73.7	72.5	62.7	67.7	51.4	48.6	62.8
	Lexical Similarity	58.3	64.0	59.7	62.7	77.3	76.9	66.5
Lat.	EigenScore	66.0	69.1	62.5	67.0	76.9	77.5	69.8
	CCS	56.9	50.3	51.7	58.2	54.2	53.2	54.1
	TSV	75.6	74.7	58.7	63.0	62.3	56.8	65.2
Traj.	TraceDet [†]	78.1	86.7	75.1	76.0	84.7	84.1	80.8
	DynHD [†]	87.3	84.4	80.1	85.6	83.5	84.6	84.3
	OSCAR (EM)	82.6	83.1	77.4	79.8	84.2	83.7	81.8
	OSCAR full pipeline (Judge)	89.5	86.3	81.9	87.1	84.3	85.1	85.7

[†] = requires trained classifier on hallucination labels.

[†] = requires trained classifier on hallucination labels.

Table 15: **Per-dataset AUROC(%) on two DLMs.** Best **bolded**, second underlined. **Green** = OSCAR detection signal under EM labels. **Blue** = OSCAR *full pipeline* (post-correction) under LLM-as-Judge labels; this is not a detector-to-detector comparison—see Table 1 for the unified detection-only protocol.

Method	TrivQA	HotQA	Category
Perplexity	52.1	50.8	Output
LN-Entropy	56.3	53.4	Output
Semantic Entropy	54.7	52.1	Output
Lexical Sim.	61.8	58.7	Output
EigenScore	65.4	61.2	Latent
TSV	58.2	56.8	Latent
OSCAR (EM)	80.3	71.5	Cross-chain
OSCAR full pipeline (Judge)	89.7	86.7	Cross-chain

Table 16: Fair-baseline AUROC (%) under unified re-evaluation (LLaDA-8B).

Variant (64 forward passes)	AUROC	Triv+Hot F1
Independent generations, standard confidence order	57.7	54.9
Reveal-order chains, confidence statistic	76.8	58.7
OSCAR (reveal-order + cross-chain entropy)	82.7	60.8

Table 17: **Compute-matched reveal-order ablation** (LLaDA-8B); all variants use the same total forward passes.

Method	TrivQA	HotQA	Avg
SelfCheck (overlap)	59.3	56.1	57.7
SelfCheck (entropy)	54.8	52.4	53.6
OSCAR (EM)	80.3	71.5	76.4
OSCAR full pipeline (Judge)	89.7	86.7	86.5

Table 18: SelfCheckGPT-DLM vs. OSCAR (LLaDA-8B, $N=8$).

Label binarization. Dropping the PARTIAL label shifts AUROC by ≤ 0.4 points with ordering preserved.

Human-labeled subset. Two NLP-trained annotators, blind to which outputs were produced by OSCAR, labeled $N=200$ examples stratified to oversample EM-vs-judge disagreements (the hardest cases for the judge). Inter-annotator agreement is $\kappa=0.71$, confirming the task is well-defined. GPT-4o $\leftrightarrow$ human agreement is $\kappa=0.76$, exceeding EM $\leftrightarrow$ human $\kappa=0.61$: the judge tracks human semantic judgment more closely than exact match does. Under human labels, OSCAR $\leftrightarrow$ DynHD remains a statistical tie ($\Delta=+0.4$, n.s.), confirming that the corrected claim—matched detection without supervision—holds regardless of evaluation label.

Dataset	AUROC	95% CI Low	High
TriviaQA (full pipeline, Judge)	89.7	86.8	92.3
HotpotQA (full pipeline, Judge)	86.7	83.2	89.9
CommQA (full pipeline, Judge)	83.4	80.1	86.5
Average (detection, EM)	76.4	73.8	79.1

Table 19: Bootstrap 95% CIs (1,000 resamples, LLaDA-8B). Full-pipeline CIs reflect post-correction outputs; the detector-to-detector comparison under the unified protocol is a statistical tie with DynHD (§5.1).

E Implementation Details

LLM-as-Judge prompt. Applied uniformly to all methods, including TraceDet and DynHD, via GPT-4o:

You are evaluating whether an AI-generated answer is factually correct. The answer does not need to match the reference exactly---it should be semantically equivalent and contain the key facts. Rate as CORRECT if the core factual content matches, INCORRECT if it contains factual errors, PARTIAL if partially correct. Respond with only the label.