Hallucination Detection in Diffusion Language Models via Sparse Token Informativeness

Anonymous ACL submission

Abstract

Diffusion large language models (DLMs) have recently emerged as a promising alternative to autoregressive LLMs, with recent systems scaled to as many as 100B-parameter. Yet hallucination in DLMs remains underexplored: we still lack a interpretable understanding of where factual errors emerge during remasking, undermining their reliability in real-world applications. Existing DLM-specific detectors recover richer denoising evidence by training supervised models over selected sub-traces or structured token-interaction graphs. We identify an *Early Factual Commitment* phenomenon in DLMs: factual responses tend to become distinguishable before generation is complete, as decisive token commitments reduce uncertainty early in the denoising trajectory. Motivated by this observation, we separate each intermediate state into two views: the already committed stable context and the remaining masked positions that encode predictive uncertainty. For each token commitment, we measure its predictive information gain by tracking how it changes uncertainty over these two views. Extensive experiments show that sparse commit-level uncertainty signals provide strong hallucination detection while requiring no auxiliary verifier, repeated sampling, dense trajectory modeling, or learned detector architecture.

1 Introduction

Large Language Diffusion Models (DLMs) (Nie et al., 2025) have recently emerged as a compelling alternative to autoregressive large language models (AR-LLMs) (Radford et al., 2019; Brown et al., 2020; Vaswani et al., 2023). Unlike AR-LLMs, which generate text sequentially via next-token prediction, DLMs iteratively predict and remask entire sequences using bidirectional attention, enabling parallel decoding and more flexible generation. Recent large-scale implementations (Bie et al., 2025) show that DLMs scale to the 100-billion-parameter regime and achieve performance comparable to similarly sized AR-LLMs across diverse tasks.

Despite these advances, DLMs remain susceptible to hallucination: the generation of linguistically plausible yet factually incorrect content (Huang et al., 2025). Many hallucination detection methods developed for AR-LLMs are ill-suited for DLMs due to fundamental architectural and generative differences (Chang et al., 2025). Existing AR-LLM detectors, such as token-level uncertainty (Fadeeva et al., 2024) or probes over static hidden representations (Azaria and Mitchell, 2023), are designed around left-to-right generation, where each output position has a single commit event conditioned on a fixed prefix. In contrast, DLMs generates through an iterative bidirectional remasking process, in which tokens may be predicted multiple times before commitment. Recent DLM-specific hallucination detectors therefore leverage richer trajectory-level dynamics, such as selected subtraces (Chang et al., 2025) or temporal token interactions (Hemmat et al., 2026). While effective, these approaches typically require additional filtering procedures, specialized architectural components, and supervised detector training. This motivates the need for a lightweight and training-free hallucination detector that can effectively leverage the remasking dynamics of DLMs.

In this work, we first identify an *Early Factual Commitment* phenomenon in DLMs: factual responses often become distinguishable before generation is complete, as decisive answer tokens commit earlier in the remasking trajectory than hallucinated responses. However, naive entropy aggregation fails to capture this phenomenon because it averages over the entire sequence and conflates uncertainty from unresolved masked tokens with already committed stable tokens.

Motivated by this observation, we decompose the remasking trajectory into two complementary views: the masked region, representing unresolved

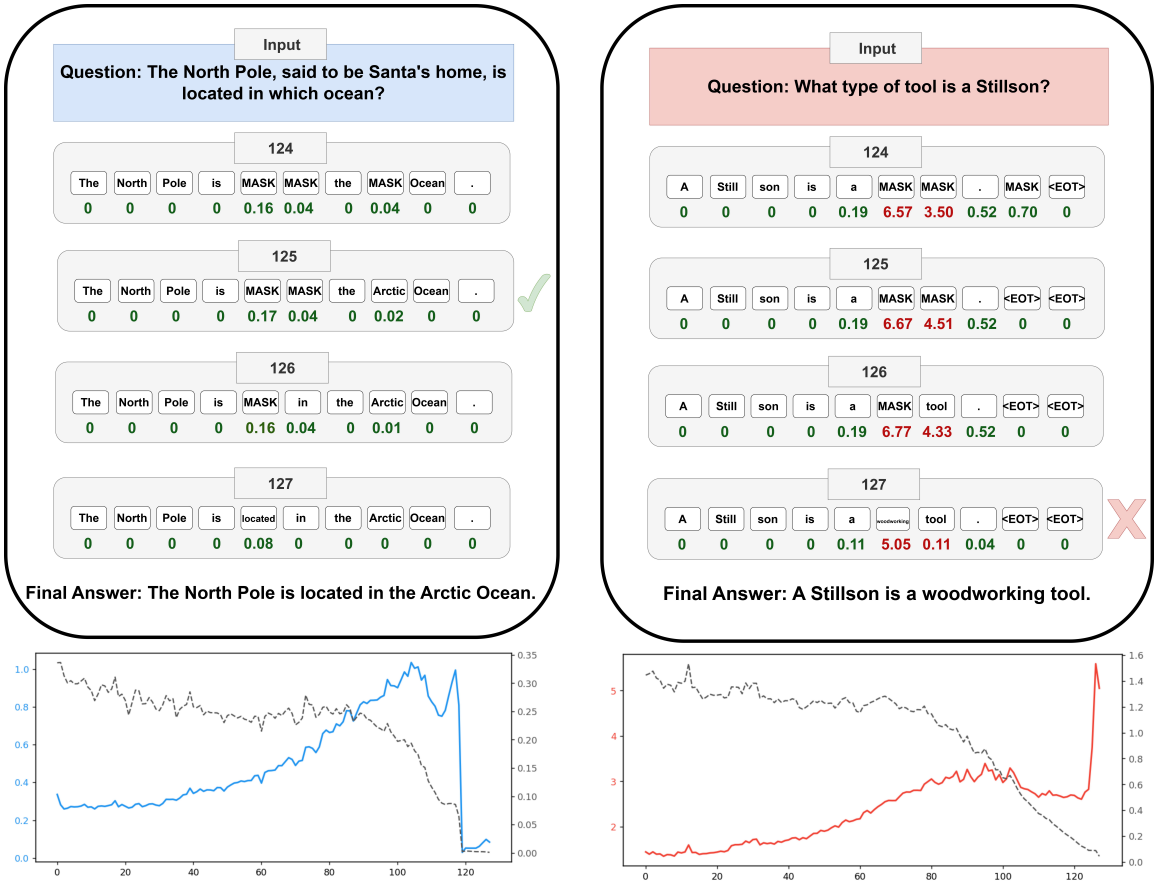


Figure 1: Illustration of token-level trajectory behavior in factual versus hallucinated generation. The top plots show the entropy of masked tokens across remasking steps. The bottom plots show the earliest step at which the partially generated response becomes identifiable as factual or hallucinated. Factual responses often become identifiable earlier in the remasking trajectory, while hallucinated responses remain uncertain until later stages. The y-axes show aggregated entropy over the MASK tokens and generated tokens, respectively. The x-axis shows remasking steps.

content, and the generated region, representing committed context. Since factual generations resolve key answer tokens earlier, the uncertainty of the remaining masked region decreases earlier and more sharply than in hallucinated generations, making masked-region entropy a natural signal for Early Factual Commitment. To further isolate decisive token commitments, we introduce a token-level *Informativeness* score that measures how strongly each commit event reshapes masked-region uncertainty. Using these top-5% commit-level signals, we construct a lightweight hallucination detector that requires no auxiliary verifier, repeated sampling, or dense trajectory modeling. Experiments on TriviaQA and HotpotQA with LLaDA-8B-Instruct and Dream-7B-Instruct show that these signals provide strong response-level hallucination detection. Our Late Mask Entropy variant reaches up to 85.03 AUROC and remains above 75.39 AUROC across all model–dataset–step settings, improving over the strongest prior training-free baseline by up to 13.77 points. Moreover, our Informativeness-based token selection is more robust to remasking policy: under random remasking, fixed late-step entropy drops to near-chance performance, while key-commit selection maintains approximately 70–74 AUROC. These results suggest that much of the hallucination evidence in DLMs can be recovered from sparse masked-region commit dynamics rather than learned dense trajectory models.

Our key contributions are summarized as follows:

- We identify Early Factual Commitment: factual DLM responses become identifiable earlier in the remasking process than hallucinated responses.
- We propose *Late Mask Entropy*, a simple de-

tector based on the entropy of still-masked positions near the end of remasking, capturing cases where important factual content remains unresolved until late generation.

- We combine late masked-region entropy with Informativeness-based key-commit selection into a lightweight hallucination detector, and further show that generated-region Consistency becomes more informative in token-editing DLMs such as LLaDA-2.1-mini.
- We show that these sparse masked-region signals achieve strong response-level hallucination detection while requiring no auxiliary verifier, repeated sampling, dense temporal graph modeling, or learned detector architecture.

2 Related Work

Diffusion Language Models. Diffusion models were first developed for continuous domains (Ho et al., 2020; Song et al., 2022) and later extended to discrete text generation (Sun et al., 2023; Li et al., 2022). Masked diffusion language models generate by iteratively predicting and remasking tokens in a partially revealed sequence, using bidirectional attention to condition on the full context. Recent large-scale DLMs such as LLaDA (Nie et al., 2025; Bie et al., 2025) and Dream (Ye et al., 2025) achieve competitive performance with autoregressive LLMs while offering greater flexibility in token commitment order through remasking strategies. Despite this progress, DLMs remain susceptible to hallucination, limiting their reliability in factual and high-stakes applications.

Hallucination Detection. Hallucination detection has been widely studied for autoregressive LLMs using prompt-based self-evaluation, token-likelihood uncertainty, consistency across sampled outputs, and probes over internal representations (Kadavath et al., 2022; Fadeeva et al., 2024; Manakul et al., 2023; Farquhar et al., 2024; Azaria and Mitchell, 2023). These methods are designed around left-to-right generation, where each token is produced once conditioned on a fixed prefix. DLMs instead generate through an iterative bidirectional denoising process in which tokens may be predicted, remasked, and committed in a non-left-to-right order. Recent DLM-specific detectors exploit this trajectory with learned components: TraceDet (Chang et al., 2025) trains a sub-trace selector and classifier, while TDGNet (Hemmat et al., 2026) learns over temporal token-interaction graphs. These approaches recover rich trajectory evidence but require supervised detector training and additional trajectory modeling. We instead study whether hallucination detection can be grounded in sparse token-level commit signals, exploiting the structural flexibility of DLM generation without learning over dense trajectory representations. Additional related work is discussed in Appendix A.

3 Method

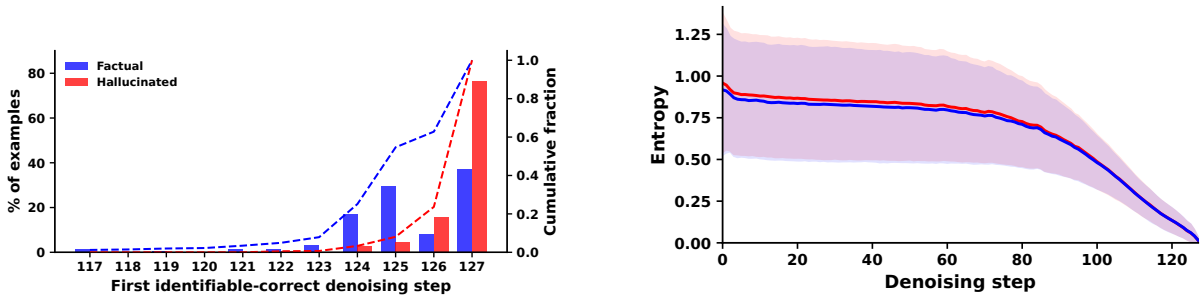
Notation. A diffusion language model begins with a fully masked response region of length L and iteratively reveals tokens over T remasking steps. We use *commit* to refer to the event in which a token position transitions from masked to a fixed vocabulary token.

At each step i, the model p_θ predicts candidate distributions for all still-masked positions, depending on the remasking strategy. In low-confidence remasking, each masked position is scored by the probability of its predicted token and the sampler commits the top-k highest-confidence positions according to the step schedule while leaving lower-confidence positions masked. Once committed, a token remains fixed for the rest of the trajectory. We denote the *masked region* $\mathcal{M}^{(i)} \subset [L]$ as the set of positions which are masked at step i and the *generated region* $\mathcal{G}^{(i)} \subset [L]$ as the set of already-committed positions, with $\mathcal{M}^{(i)} = [L] \setminus \mathcal{G}^{(i)}$.

3.1 Early Factual Commitment

In this section, we motivate our method by investigating the early commitment of correct answers in diffusion language models. Earlier works such as (Li et al., 2026) have discovered that the correct answer can be internally identified many steps before the final decoding step, both under semi-autoregressive and confidence-based remasking schedules. This raises a natural question: do factual answers become identifiable earlier in the remasking trajectory than hallucinated ones?

We conduct an empirical investigation using LLaDA-8B-Instruct (Nie et al., 2025) and subsets of two commonly used QA benchmarks: TriviaQA (Joshi et al., 2017) and HotpotQA (Yang et al., 2018). Specifically, we examine at which timestep the committed sequence can be classified as factual or hallucinated. We present the results of standard remasking strategy in Figure 2a.



(a) (b)

Figure 2: Illustration of early factual commitment on LLaDA. (a) While LLaDA tends to provide both factual and hallucinated answers in late-generation stages, factual responses and hallucinated responses differ in commit timing. (b) Naive mean entropy cannot capture the difference in commit timing between hallucinated and factual instances and cannot distinguish hallucination from factual answers.

Factual Answers Become Identifiable Earlier Than Hallucinated Answers. While both Dream and LLaDA tends to commit key answer tokens in late-generation stages, hallucinated answers and factual answers also differ by timestep which they can be identified. When committing the most confident token, only 16.7% of factual TriviaQA answers were identifiable until the final stage, while the number is 35.6% and 74.5% in the final 2 steps for hallucinated figures. When committing the top-2 most confident tokens, 91.9% of hallucinated answers were not identifiable until the final 2 steps while 40.3% of factual answers could be identified earlier in the remasking schedule. Similar trends also hold for HotpotQA answers. This reveals that LLaDA often resolve decisive answer information earlier than hallucinated generations and can be a useful signal for hallucination detection.

Mean Entropy Fails to Capture Early Factual Commitment. As discussed in previous works (Chang et al., 2025), we confirm that naive aggregation of entropy fail to capture the Early Factual Commitment phenomenon. As shown in Figure 2b, hallucinated and factual instances exhibit similar mean-entropy trajectories across early, intermediate, and late stages of generation. Although factual generations have slightly lower entropy than hallucinated generations, the difference is largely within the error range, rendering mean entropy an unreliable detection signal. This distinction is especially natural in DLMs: at any remasking step, the already committed tokens form the observed context, while the still-masked tokens represent the unresolved content to be predicted.

3.2 Predictive Uncertainty

We hypothesize that averaging token-level entropy over the full sequence fails because it conflates two different quantities: uncertainty over unresolved masked positions and the model's support for already committed tokens. Let $\mathcal{M}^{(i)}$ denote the still-masked region and $\mathcal{G}^{(i)}$ the already committed region at remasking step i. The full-sequence mean entropy can be decomposed as

$$E^{(i)}_{\text{mean}} = \frac{|\mathcal{M}^{(i)}|}{L} E^{(i)}_{\text{masked}} + \frac{|\mathcal{G}^{(i)}|}{L} E^{(i)}_{\text{gen}}. \quad (1)$$

Thus, full-sequence entropy mixes uncertainty over tokens that still need to be predicted with uncertainty over tokens that have already been fixed.

For hallucination detection, the first term is the more relevant one: $E^{(i)}_{\text{masked}}$ measures the model's predictive uncertainty (Bialek et al., 2001) over the unresolved part of the response, conditioned on the currently committed context. A low masked-region entropy means that the observed committed tokens strongly constrain the remaining masked positions; a high masked-region entropy means that important content remains unresolved.

This perspective also gives a commit-level notion of predictive information gain. Let $\mathcal{B}^{(i)}$ be the set of tokens committed at step i, and let $\mathcal{R}^{(i)} = \mathcal{M}^{(i+1)}$ be the positions that remain masked after that commit. The information gained about the remaining masked positions can be approximated by the reduction in their predictive entropy:

$$\text{IG}^{(i)} = E^{(i)}_{\mathcal{R}^{(i)}} - E^{(i+1)}_{\mathcal{R}^{(i)}}. \quad (2)$$

This quantity compares uncertainty over the same unresolved positions before and after the commit.

A positive value indicates that the newly committed token helps predict the remaining masked content, while a negative value indicates that the commit makes the remaining content less certain.

3.3 Mask Entropy and Generation Entropy

We operationalize the above decomposition by tracking uncertainty separately over the still-masked and already committed regions. We define mask entropy and generation entropy as

$$E^{(i)}_{\text{masked}} = \frac{1}{|\mathcal{M}^{(i)}|} \sum_{k \in \mathcal{M}^{(i)}} H\Big(p_\theta(x_k \mid \mathbf{x}^{(i)})\Big),$$
$$E^{(i)}_{\text{gen}} = \frac{1}{|\mathcal{G}^{(i)}|} \sum_{k \in \mathcal{G}^{(i)}} H\Big(p_\theta(x_k \mid \mathbf{x}^{(i)})\Big). \tag{3}$$

Here, $E^{(i)}_{\text{masked}}$ measures predictive uncertainty over unresolved content, while $E^{(i)}_{\text{gen}}$ measures the model's continued support for already committed content. Normalizing by region size makes the two quantities comparable across remasking steps as the masked region shrinks.

3.4 Informativeness and Consistency

Next, we define commit-level scores that measure how each commitment changes uncertainty in the remasking trajectory. We use the step index i directly: when one or more tokens are committed between $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(i+1)}$, we treat that transition as a single commit event. If no token is committed at step i, we the corresponding scores are zero.

For each non-final step i, we define two entropy-based importance scores:

$$\mathcal{I}^{(i)} = E^{(i)}_{\text{masked}} - E^{(i+1)}_{\text{masked}}, \quad \mathcal{C}^{(i)} = E^{(i)}_{\text{gen}} - E^{(i+1)}_{\text{gen}}. \tag{4}$$

Informativeness $\mathcal{I}^{(i)}$ measures how the commit affects uncertainty over the remaining masked region. A positive value indicates that uncertainty decreases after the commit, meaning that the newly committed token helps constrain unresolved positions. A negative value indicates that masked uncertainty increases after the commit. Thus, $|\mathcal{I}^{(i)}|$ captures how strongly the commit reshapes the remaining trajectory, regardless of whether the effect is helpful or harmful.

Consistency $\mathcal{C}^{(i)}$ measures how the generated region responds after the new token becomes visible. A positive value indicates that the generated region becomes more stable, while a negative value indicates that the commit destabilizes already committed content. Together, $\mathcal{I}^{(i)}$ and $\mathcal{C}^{(i)}$ characterize whether a commit helps resolve the remaining masked positions and remains compatible with the generated context.

Dataset	Steps	Last 1	Last 5	Last 10
TriviaQA	128	0.8010	**0.8686**	0.8486
HotpotQA	128	0.7589	**0.8070**	0.7939
TriviaQA	64	0.8263	**0.8521**	0.8259
HotpotQA	64	0.7906	**0.8093**	0.8046

Table 1: AUROC of hallucination detection using masked-region entropy aggregated over the final remasking steps. **Steps** denotes the remasking schedule length. **Last 1** uses $E^{(T-1)}_{\text{masked}}$, while **Last 5** and **Last 10** average masked-region entropy over the final 5 and 10 remasking steps, respectively. Results are computed on 1000 randomly sampled questions from each dataset.

3.5 Hallucination Detector

We define two hallucination detectors based on masked-region entropy: *Late Mask Entropy (LME)* and *Token Informativeness (TI)*. Both detectors use entropy statistics already produced during a single remasking run, but they differ in how they select the relevant timesteps.

Late Mask Entropy (LME). As shown in Section 1, masked-region entropy captures the dynamics of Early Factual Commitment through the rising tail of uncertainty in late remasking stages. Under confidence-based remasking, low-confidence and factuality-critical tokens are often postponed to the end, making late-stage masked entropy a strong signal. We therefore define LME as the mean masked-region entropy over the final m remasking steps:

$$S^{(m)}_{\text{LME}} = \frac{1}{m} \sum_{j=1}^{m} E^{(T-j)}_{\text{masked}}. \tag{5}$$

In our experiments, we use $m = 5$ for LME. Table 1 shows that the final-5-step average achieves the best AUROC among the late-step aggregation choices.

Token Informativeness (TI). Although LME is effective under confidence-based remasking, it relies on the assumption that key tokens occur near the final remasking steps. Under other remasking strategies, informative commits may occur at different absolute timesteps. To reduce this dependence, we define *Token Informativeness (TI)*, which selects

steps by how strongly they change masked-region uncertainty.

For each non-final remasking step $i \in \{0, \ldots, T-2\}$, we use the magnitude of Informativeness as the key-step score:

$$a_i = |\mathcal{I}^{(i)}|. \tag{6}$$

If no token is committed at step i, we set $a_i = 0$. Let $\mathrm{Top}_\rho(a)$ denote the indices of the top ρ fraction of non-final steps ranked by a_i, rounding up to at least one step. We aggregate masked-region entropy over these selected steps and combine it with final masked-region entropy:

$$\begin{aligned} S_{\mathrm{key}}^{(\rho)} &= \frac{1}{|\,\mathrm{Top}_\rho(a)|} \sum_{i \in \mathrm{Top}_\rho(a)} E_{\mathrm{masked}}^{(i)}, \\ S_{\mathrm{TI}} &= w_1 S_{\mathrm{key}}^{(\rho)} + w_2 E_{\mathrm{masked}}^{(T-1)}. \end{aligned} \tag{7}$$

Here, $S_{\mathrm{key}}^{(\rho)}$ captures sparse high-impact commit events. Since $\mathcal{I}^{(i)}$ is undefined at the final non-empty masked step, TI excludes that step from key-step selection and includes $E_{\mathrm{masked}}^{(T-1)}$ separately as the late-resolution signal. Unless otherwise stated, we use $\rho = 0.05$. The weights (w_1, w_2) and the decision threshold are selected on a 200-sample calibration set. TI requires only a single remasking run and does not train a learned detector architecture.

4 Experiments

4.1 Setup

Datasets. We evaluate hallucination detection on two factuality-oriented QA benchmarks: TriviaQA (Joshi et al., 2017), an open-domain factoid QA dataset, and HotpotQA (Yang et al., 2018), a multi-hop QA dataset requiring evidence aggregation across supporting contexts. These datasets test complementary forms of factual generation: single-hop entity-centric answers in TriviaQA and compositional multi-hop answers in HotpotQA.

For each dataset, we construct a pool of questions and generate responses from each D-LLM. Each response is labeled as factual or hallucinated by comparing the generated answer against the reference answer. For methods that require calibration, we use a small 200-sample calibration set and evaluate on the remaining held-out examples. Unless otherwise stated, all reported AUROC scores are computed on the held-out split.

Models. We evaluate on two representative diffusion language model backbones: LLaDA-8B-Instruct (Nie et al., 2025) and Dream-7B-Instruct (Ye et al., 2025). We evaluate each model under 64-step and 128-step remasking schedules to test whether the proposed signals remain robust across generation lengths. For Dream, we filter out the first 10 remasking steps due to known problem with mode collapse.

Response labeling. For open-ended QA, we use GPT-4.1-nano (OpenAI et al., 2024) as an automatic evaluator to assign factuality labels for generated responses. The evaluator is given the question, the model response, and the gold answer, and returns a binary judgment indicating whether the response is semantically correct with respect to the reference. This evaluator is used only for offline evaluation labeling and is not used by any hallucination detector at inference time. To assess the reliability of these labels, we conduct a small manual verification study on a small subset of TriviaQA and HotpotQA samples and find high agreement between the automatic judgments and manual annotations. We further discuss our verification method and include examples in Appendix section B.

Baselines. We compare against both training-free and training-based hallucination detection baselines. Training-free baselines include Perplexity (Ren et al., 2023), Length-Normalized Entropy (LN-Entropy) (Malinin and Gales, 2021), Semantic Entropy (Kuhn et al., 2023), Lexical Similarity (Lin et al., 2024), and EigenScore (Chen et al., 2024). We also include simple trajectory-level entropy baselines, including full-sequence mean entropy and late-step mask entropy. Training-based baselines include methods that learn additional detector components or trajectory representations, such as TSV (Park et al., 2025) and TraceDet (Chang et al., 2025).

Metric. Following prior work, we report AUROC as the primary metric for hallucination detection. Higher AUROC indicates better separation between factual and hallucinated responses.

4.2 Main Results

Table 2 reports the main hallucination detection results across datasets, models, and remasking lengths. Our proposed masked-region signals are consistently strong: LME remains above 75.39 AUROC across all settings, improving over the

Table 2: AUROC(%) comparison of hallucination detection methods on D-LLMs across TriviaQA and HotpotQA with 128-step and 64-step remasking schedules. Methods are grouped by whether they require training an auxiliary detector or probe. SS denotes whether the method uses a single sampling. **Ours-LME** uses Late Mask Entropy, the mean mask entropy over the final five remasking steps. **Ours-TI** uses Informativeness-based key-commit selection and calibrates the scalar weights and threshold on 200 samples. The highest score is **bolded** and the second highest is underlined.

Model	Method	SS	TriviaQA		HotpotQA	
			128	64	128	64
	Training-Free					
LLaDA-8B-Instruct	Perplexity	✗	57.63	47.40	65.40	52.00
	LN-Entropy	✓	59.32	62.72	58.16	69.31
	Semantic Entropy	✗	64.92	76.90	70.33	81.10
	Lexical Similarity	✗	66.81	69.80	70.96	76.05
	EigenScore	✗	57.00	61.98	64.38	70.33
	Ours-LME	✓	**78.98**	**78.53**	**84.73**	**85.03**
	Training-Based					
	TSV	✓	54.68	58.88	60.07	67.17
	TraceDet	✓	66.20	71.70	76.00	75.20
	Ours-TI	✓	78.33	77.61	83.41	82.73
	Training-Free					
Dream-7B-Instruct	Perplexity	✗	69.15	68.77	63.23	64.31
	LN-Entropy	✓	55.85	54.44	54.53	62.82
	Semantic Entropy	✗	64.05	66.50	69.65	74.04
	Lexical Similarity	✗	64.91	65.25	70.35	70.54
	EigenScore	✗	63.81	68.94	68.64	76.97
	Ours-LME	✓	**76.37**	**75.39**	**81.44**	**80.86**
	Training-Based					
	TSV	✓	73.16	73.81	75.44	76.38
	TraceDet	✓	70.40	72.60	72.90	75.50
	Ours-TI	✓	73.17	74.51	79.19	77.79

Table 3: AUROC(%) for LLaDA-8B-Instruct under low-confidence and random remasking. Error bars are 95% confidence intervals over 20 trials.

Dataset	Steps	Remasking Strategy	**Ours-TI** Top-5%	**Ours-TI** Top-25%	**Ours-LME** Last-5
TriviaQA	64	Low-confidence	78.95 ± 0.50	77.94 ± 0.57	80.22 ± 0.19
		Random	70.28 ± 0.39	71.43 ± 0.41	53.05 ± 0.15
	128	Low-confidence	79.21 ± 0.23	77.61 ± 0.40	80.61 ± 0.18
		Random	72.12 ± 0.23	73.68 ± 0.27	53.47 ± 0.17
HotpotQA	64	Low-confidence	84.10 ± 0.19	83.66 ± 0.19	86.17 ± 0.11
		Random	72.61 ± 0.59	73.79 ± 0.57	55.81 ± 0.16
	128	Low-confidence	83.57 ± 0.16	82.62 ± 0.17	85.26 ± 0.14
		Random	72.52 ± 0.30	73.77 ± 0.34	55.11 ± 0.24

strongest prior training-free baseline by up to 13 points. TI, which selects key commits by Informativeness, remains competitive while avoiding fixed late-step aggregation.

These results suggest that much of the useful hallucination evidence in DLMs is exposed by sparse masked-region uncertainty, rather than requiring learned trajectory models over selected sub-traces or dense token-interaction graphs.

Robustness to Remasking Strategy. We next test whether our signals depend on the remasking policy. We compare LME, which uses the final five remasking steps, with TI, which selects the top Informativeness-ranked commit events independent of their absolute timestep. Results are reported in Table 3.

As shown in Table 3, LME is highly effective under low-confidence remasking, where the final masked positions are explicitly enriched for uncer-

tain tokens. However, this signal is strongly tied to the remasking policy. Under random remasking, the method drops to near-chance performance, with AUROC around 53–56%.

In contrast, TI remains substantially more robust. Although random remasking reduces performance, the Top-5% and Top-25% variants still maintain approximately 70–74% AUROC. This shows that Informativeness-based selection is not merely exploiting the fact that low-confidence remasking postpones difficult tokens to the end. Instead, it identifies influential commit events by their entropy-dynamic effect, recovering sparse token-level evidence even when key commits are no longer concentrated in the final remasking steps.

Computation. Our detector adds negligible overhead over clean single-run inference: for 100 TriviaQA examples with LLaDA-8B-Instruct, runtime increases from 392.4s to 394.8s. In contrast, sampling-based baselines requiring four generations take roughly 1,400s. We provide full comparisons between baselines in Section C.

5 Analysis

Our main detector uses Informativeness-based key-commit selection together with final masked-region uncertainty. We further ask whether generated-region signals become useful in newer diffusion decoders. To study this, we evaluate LLaDA-2.1-mini, a recent LLaDA-family model with Token-to-Token decoding (T2T). More analysis on the entropy dynamic of LLaDA-2.1-mini and LLaDA-8B-Instruct are provided in Section D.

Variants. We compare our original detector, TI, against three controlled variants. TI selects the top 5% commit events ranked by Informativeness magnitude $|\mathcal{I}^{(s)}|$ and combines their selected-step mask entropy with final masked-region entropy. The three variants modify either the final-region signal or the commit-selection signal, respectively. Variant 1 keeps the same Informativeness-based commit selection as TI, but replaces final masked-region entropy with final generated-region entropy. Variant 2 instead selects commits using Consistency magnitude $|\mathcal{C}^{(s)}|$ and combines the selected-step score with final generated-region entropy. Variant 3 also selects commits using Consistency magnitude $|\mathcal{C}^{(s)}|$, but combines the selected-step score with final masked-region entropy.

As shown in Table 4, Informativeness remains the stronger key-commit selection signal on LLaDA-2.1-mini: TI and Variant 1 both outperform the Consistency-selected variants. At the same time, generated-region entropy becomes more informative than in the earlier D-LLMs studied in our main experiments. On TriviaQA, Variant 1 slightly outperforms TI, suggesting that generated-region uncertainty can provide useful complementary evidence under Token-to-Token decoding. These results suggest that generated-region dynamics are not inherently uninformative; their utility depends on the decoding mechanism and may increase as D-LLMs move toward more interactive token-editing. Nevertheless, Informativeness-based key-commit selection remains the strongest and most consistent signal in our main experiments. Additional LLaDA-2.1 analysis is provided in Section D.

Table 4: AUROC of LLaDA-2.1-mini detector variants on 2000-questions subsets of TriviaQA and HotpotQA. Error bars are 95% confidence intervals over 20 trials.

Method	TriviaQA	HotpotQA
TI	0.8460 ± 0.0047	$\mathbf{0.8221 \pm 0.0051}$
Variant 1	$\mathbf{0.8564 \pm 0.0028}$	0.8172 ± 0.0046
Variant 2	0.8224 ± 0.0033	0.7812 ± 0.0053
Variant 3	0.8119 ± 0.0030	0.7837 ± 0.0047

6 Conclusion

We studied hallucination detection in diffusion language models through denoising-time uncertainty dynamics. We identify *Early Factual Commitment*: factual responses often become identifiable earlier than hallucinated responses. Motivated by this observation, we decompose the denoising state into still-masked and already committed regions, showing that masked-region entropy provides a strong response-level hallucination signal. Late Mask Entropy captures unresolved factual content near the end of denoising, while Token Informativeness identifies high-impact commit events whose entropy effects are not tied to fixed timesteps. Together, these signals enable efficient hallucination detection on DLMs without auxiliary verifiers, repeated sampling, or learned detector architectures. Our LLaDA-2.1-mini analysis further suggests that generated-region dynamics may become useful in more interactive token-editing D-LLMs.

7 Limitations

Our findings point to several directions that require further study. First, the effectiveness of

masked-region entropy depends on the remasking strategy. Late Mask Entropy is highly effective under low-confidence remasking, where uncertain tokens are naturally postponed to the final denoising stages. However, under random remasking, this late-step signal drops substantially. Although Informativeness-based key-commit selection is more robust, its performance also decreases under random remasking. Further work is needed to develop remasking-aware normalization or selection strategies that preserve sparse token-level signals across a broader range of decoding policies.

Second, our experiments focus on QA-style factuality benchmarks and two primary D-LLM backbones, LLaDA-8B-Instruct and Dream-7B-Instruct. These settings are useful for controlled response-level hallucination evaluation, but they do not cover the full range of D-LLM applications. Future work should study whether Early Factual Commitment and sparse masked-region signals generalize to long-form generation, retrieval-augmented generation, dialogue, summarization, and newer token-editing diffusion models.

Finally, our detector uses a small held-out calibration set to select scalar weights and thresholds. Although this does not train a learned detector architecture, fully calibration-free variants would be valuable for deployment in low-resource or distribution-shifted settings. Our additional analysis on LLaDA-2.1-mini also suggests that generated-region dynamics may become more informative in newer decoding schemes, motivating further study of Consistency and generated-region uncertainty as complementary hallucination signals.

Acknowledgments

The authors used Claude Code to assist with implementing some experiments and ChatGPT to support grammar correction and writing refinement. These tools were not used to generate scientific claims or replace author judgment. The authors reviewed, edited all LLM outputs and assume full responsibility for the content of the manuscript.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

William Bialek, Ilya Nemenman, and Naftali Tishby. 2001. Predictability, complexity and learning. *Preprint*, arXiv:physics/0007070.

Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jiaqi Hu, Zenan Huang, Zhenzhong Lan, Chengxi Li, Chongxuan Li, Jianguo Li, Zehuan Li, Huabin Liu, Lin Liu, Guoshan Lu, Xiaocheng Lu, Yuxin Ma, and 12 others. 2025. Llada2.0: Scaling up diffusion language models to 100b. *Preprint*, arXiv:2512.15745.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

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

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9367–9385.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Preprint*, arXiv:2006.11239.

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

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Preprint*, arXiv:1705.03551.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

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

Pengxiang Li, Yefan Zhou, Dilxat Muhtar, Lu Yin, Shilin Yan, Li Shen, Soroush Vosoughi, and Shiwei Liu. 2026. Diffusion language models know the answer before decoding. *Preprint*, arXiv:2508.19982.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Preprint*, arXiv:2205.14217.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. *Preprint*, arXiv:2002.07650.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *Preprint*, arXiv:2502.09992.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. 2025. Steer llm latents for hallucination detection. *Preprint*, arXiv:2503.01917.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. 2023. Out-of-distribution detection and selective generation for conditional language models. *Preprint*, arXiv:2209.15558.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2022. Denoising diffusion implicit models. *Preprint*, arXiv:2010.02502.

Haoran Sun, Lijun Yu, Bo Dai, Dale Schuurmans, and Hanjun Dai. 2023. Score-based continuous-time discrete diffusion models. *Preprint*, arXiv:2211.16750.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

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

Appendix

Contents

A Additional Related Work 11

B LLM Labeling 12

C Runtime 12

D Additional Analysis on LLaDA-2.1-Mini 12

E Additional Analysis on Early Factual Commitment 13

F Token Selection via Informativeness Score 13

A Additional Related Work

Hallucination Detection in AR-LLMs. Hallucination detection has been extensively studied in the autoregressive setting. Prompt-based methods elicit a model's uncertainty through targeted prompting or self-evaluation (Kadavath et al., 2022). Likelihood-based methods use token probabilities or sequence likelihoods as confidence signals for factuality (Fadeeva et al., 2024). Consistency-based methods compare independently sampled outputs, treating disagreement across responses as evidence of hallucination (Manakul et al., 2023). Latent-based methods instead probe hidden states or internal representations to predict whether a generated response is factual (Azaria and Mitchell, 2023; Chen et al., 2024). These approaches are naturally aligned with autoregressive generation, where each output position is selected once conditioned on a fixed prefix. DLMs, on the other hand, have a different structure: a position may receive multiple candidate predictions while still masked, and it becomes visible to other positions only when it is committed. This motivates DLM-specific signals based on commit-time uncertainty rather than directly reusing features designed for left-to-right decoding.

Hallucination Detection in DLMs. DLM hallucination detection is challenging because generation unfolds over an iterative denoising trajectory rather than a single left-to-right pass. Recent methods exploit this trajectory with learned detectors: TraceDet (Chang et al., 2025) trains a sub-trace selector and classifier, while TDGNet (Hemmat et al., 2026) learns over temporal token-interaction graphs. These approaches recover rich trajectory evidence but require supervised detector training and additional trajectory modeling. We instead study whether sparse token-level commit signals are sufficient for hallucination detection.

Uncertainty Quantification. Uncertainty quantification for language generation has been studied through output probabilities, semantic variability, sampling consistency, and internal representations. Early works, such as (Malinin and Gales, 2021), study uncertainty estimation for autoregressive structured prediction, while semantic uncertainty methods measure uncertainty over meaning rather than surface forms (Kuhn et al., 2023; Farquhar et al., 2024). Black-box approaches estimate reliability from sampled generations or output vari-

ability (Lin et al., 2024; Ren et al., 2023). Our method differs in that it uses uncertainty already produced during a single DLM remasking run, focusing on masked-region entropy and commit-level Informativeness rather than repeated sampling or trained uncertainty probes.

B LLM Labeling

For open-ended QA, exact keyword matching can underestimate factual correctness: a generated answer may be semantically correct even when it does not exactly match the reference string, and conversely may contain a reference keyword while still giving an incorrect answer. We therefore use ChatGPT-4.1-nano as an automatic evaluator for response-level factuality. The evaluator is given the question, model response, and gold answer, and returns a binary judgment of whether the response is semantically correct. This evaluator is used only for offline labeling and is not used by any hallucination detector at inference time.

To assess label reliability, we manually verify a small subset consisting of the first 100 examples from TriviaQA and the first 100 examples from HotpotQA. As shown in Table 5, GPT-4.1-nano agrees with manual labels on approximately 98% of inspected examples, outperforming string matching by up to 8.1 percentage points. In particular, the LLM evaluator better handles paraphrases and semantically equivalent answers than keyword matching, as shown by the examples in Table 6.

Table 5: Agreement with manual factuality labels on manually inspected subsets.

Dataset	Method	Agreement
TriviaQA	GPT-4.1-nano	98.0%
	String match	89.9%
HotpotQA	GPT-4.1-nano	98.0%
	String match	92.0%

C Runtime

We report inference-time runtime in Table 7. Because our detector reuses the logits already produced during the standard generation trace, it adds negligible overhead over clean inference: runtime increases from 392.4s to 394.8s for 100 TriviaQA examples. In contrast, sampling-based baselines that require $n = 4$ generations take roughly $1{,}400$s, making them about $3.6\times$ slower. This shows that our method preserves the efficiency of single-run inference while avoiding the repeated sampling cost of output-variability methods.

Table 6: Representative examples where semantic LLM-based labeling agrees with human judgment but exact string matching fails.

Example 1 [HotpotQA]
Q: The inventor of Mayer's reagent and one of the founders of thermodynamics, Julius von Mayer was the first to enunciate what phrase, known as the first law of thermodynamics?
A: Conservation of energy.
Gold: energy can be neither created nor destroyed
Aliases: None
GPT: True **String Match:** False

Example 2 [TriviaQA]
Q: Who has been Israel's only female Prime Minister?
A: Gold Meir.
Gold: Golda Meir
Aliases: None **GPT:** True **String Match:** False

Table 7: Estimated wall-clock runtime for 100 TriviaQA examples with LLaDA-8B-Instruct using 128 denoising steps. Sampling-based baselines use $n = 4$ generations.

Method	Wall Time (s)
Clean inference	392.4
Ours entropy trace	394.8
LN-Entropy	394.8
TSV	387.6
Perplexity (MC $= 128$)	433.8
Lexical Similarity ($n = 4$)	1411.8
EigenScore ($n = 4$)	1407.6
Semantic Entropy ($n = 4$)	1414.8

D Additional Analysis on LLaDA-2.1-Mini

In this section, we provide additional analysis of LLaDA-2.1-mini, extending the results in Section 5. Our main experiments show that masked-region signals are the most reliable indicators for LLaDA-8B-Instruct and Dream-7B-Instruct. However, LLaDA-2.1-mini introduces Token-to-Token decoding (T2T), which allows more interactive token-editing behavior during generation. This makes it a useful test case for studying whether generated-region entropy and Consistency become more informative in newer D-LLM decoding schemes.

As shown in Figure 3a, LLaDA-2.1-mini exhibits substantially richer generated-region entropy dynamics than LLaDA-8B-Instruct. For LLaDA-8B-Instruct, generated-region entropy largely trends toward zero, with only local fluctuations. In

contrast, the generated-region entropy of LLaDA-2.1-mini changes noticeably over the remasking trajectory and shows different patterns for factual and hallucinated responses. This suggests that already committed tokens are more actively re-evaluated under Token-to-Token decoding. More broadly, generated-region uncertainty is not inherently uninformative; its utility depends on the model architecture and decoding mechanism.

Table 4 and Figure 3b further show that Informativeness remains the stronger key-commit selection signal: TI and Variant 1 outperform the Consistency-based variants. At the same time, the gap between final masked-region and final generated-region entropy becomes smaller than in earlier D-LLMs. In particular, the generated-region variant is competitive with TI, indicating that generated-region dynamics can provide complementary hallucination evidence under Token-to-Token decoding.

Overall, these results support a broader interpretation of our masked/generated-region decomposition. For standard mask-prediction D-LLMs, masked-region uncertainty remains the primary signal because uncertainty is concentrated in unresolved positions. For more interactive token-editing decoders such as LLaDA-2.1-mini, generated-region entropy and Consistency may become increasingly useful because already committed tokens are more actively re-evaluated during generation. These findings suggest that generated-region entropy and Consistency may provide useful signals for future research on hallucination detection in interactive token-editing DLMs.

E Additional Analysis on Early Factual Commitment

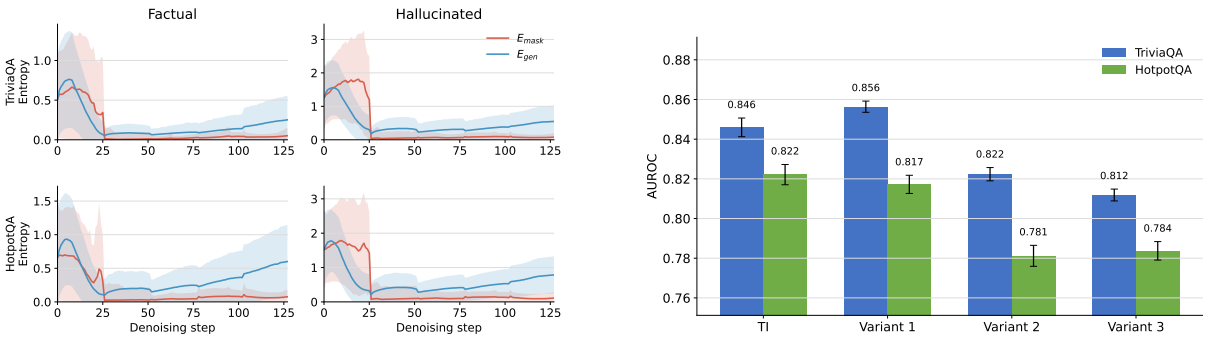
To validate the Early Factual Commitment phenomenon and the reliability of our LLM-based judge, we manually inspect the first 100 examples from both TriviaQA and HotpotQA. We find high agreement between the judge and manual assessment. Representative examples are shown in Table 8.

The examples show that a response can become factually identifiable before the entire sequence is finalized: once the answer-bearing token or phrase is committed, remaining masked positions often correspond to punctuation, padding, or non-essential continuation tokens. We hypothesize that this occurs because confidence-based remasking commits strongly constrained factual tokens earlier. Once these decisive tokens become visible, they stabilize the remaining denoising trajectory. Hallucinated responses, in contrast, often postpone the key entity, date, relation, or attribute until late denoising, or resolve it only after easier tokens have already been committed.

F Token Selection via Informativeness Score

In this section, we analyze which types of tokens are selected by the Informativeness-based key-commit score. For each generated response, we rank commit events by Informativeness magnitude $|\mathcal{I}^{(i)}|$ and retain the top 5% of commits. Since a committed token may correspond to a subword or a partial number, we map each selected token back to its surrounding surface form in the final generated answer before assigning a semantic category. For example, if the selected token is `1` in the final string `1996`, we categorize the selected commit as a year rather than as an isolated digit.

As shown in Table 9, high-Informativeness commits in factual responses are frequently content-bearing tokens, including places, person names, years, and other answer-relevant content words. This supports the interpretation that Informativeness identifies commits that substantially affect the unresolved masked region. In hallucinated responses, however, a much larger fraction of selected commits correspond to structural tokens. These results suggest that Informativeness often selects semantically meaningful commit events, but they should be interpreted as an exploratory analysis rather than a definitive taxonomy of hallucination-relevant tokens. Nevertheless, further study would be useful to determine whether these category patterns generalize across broader datasets, long-form generations, and newer DLM decoding strategies.



(a) (b)

Figure 3: Entropy dynamic and performance of our hallucination detector on LLaDA-2.1-Mini. (Left) Mean mask entropy E_{mask} and generation entropy E_{gen} over denoising steps for factual and hallucinated responses on TriviaQA and HotpotQA. (Right) Heldout AUROC for TI and three controlled variants on TriviaQA and HotpotQA (LLaDA-2.1-mini). Error bars show ±1 std over 20 random train/test splits.

Table 8: Examples of early identifiable answers during denoising. The generated answer column shows the committed visible content at the step where the answer first became identifiable.

Question	**Gold Answer**	**Identifiable Step**	**Generated Answer at the Time Step**
Which of the Brontë sisters wrote *The Tenant of Wildfell Hall* in 1848?	Acton Bell	113	`Anne Brontë[M][M][M][M]...`
Which legendary English outlaw "stole from the rich to give to the poor"?	Robin of Locksley	113	`Robin Hood[M][M][M][M]...`
Who wrote "To Kill a Mockingbird"?	Harper Lee	116	`Harper[M][M][M][M]...`
What is the most populous city in Alaska?	Anchorage	117	`Anchorage[M][M][M][M]...`
What was the native language of Albert Einstein?	German	119	`German[M][M][M][M]...`
The Nobel Peace Prize is the only Nobel prize not awarded in Stockholm. In what city is it awarded?	Oslo	121	`Oslo[M][M][M][M]...`
What Russian empress had affairs with many men?	Russian roots of Catherine the Great	122	`Catherine[M][M][M][M]...`
The Statue of Liberty in New York Harbour was a gift in 1886 to the USA by whom?	La Republique francaise	123	`France[M][M][M][M]...`
Prior to the introduction of the Euro, which coin was equal to one-hundredth of a French Franc?	Centime	124	`Cent[M][M][M]...`
In the novel *Jane Eyre* by Charlotte Brontë, what is the name of Mr. Rochester's home?	Thornfield Hall	124	`Thornfield[M][M][M]...`
Before 1966, the Republic of Botswana was a protectorate of the British Government and was known as what?	Bechuanaland protectorate	125	`Bechuanaland[M][M]...`
What is the name of the force which keeps the planets in orbit around the sun?	Gravitational physics	125	`Gravitational force[M][M]...`
Known as the Gentle Giant, name the breed of dog that is recognized as the tallest of all.	Irish wolf hound	126	`[M] Dane...`

	TriviaQA		**HotpotQA**	
Category	**Factual**	**Halluc.**	**Factual**	**Halluc.**
Place	12.8%	5.4%	13.2%	5.0%
Person Name	15.0%	15.3%	17.1%	16.9%
Time (Year)	1.7%	2.6%	3.5%	9.3%
Other content word	28.8%	12.3%	27.4%	12.5%
Filler	5.8%	5.2%	8.3%	6.1%
Structural Tokens	13.0%	38.0%	14.2%	32.6%
Punctuation	20.2%	20.3%	12.4%	12.7%

Table 9: Distribution of selected key commits across semantic categories. We select the top 5% of commit events by Informativeness magnitude $|\mathcal{I}^{(i)}|$ and categorize each selected token according to its surrounding surface form in the final generated answer. Percentages may not sum exactly to 100 due to rounding and uncategorized tokens.