HALLUCINATION DETECTION IN DIFFUSION LANGUAGE MODELS VIA SPARSE TOKEN INFORMATIVENESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion large language models (DLMs) have recently emerged as a promising alternative to autoregressive LLMs, yet hallucination in DLMs remains underexplored, with limited understanding of how factual errors emerge during the remasking process. Existing DLM-specific detectors exploit denoising trajectories through supervised models over selected sub-traces or structured token interactions, introducing additional training and trajectory modeling. In this work, we first identify an early factual commitment phenomenon in LLaDA, where factual responses tend to reveal answer-defining information earlier than hallucinated responses. This suggests that factuality-related evidence is concentrated in particular commitment events rather than distributed uniformly across the denoising trajectory. To capture this structure, we decompose each intermediate state into masked and generated regions and track predictive uncertainty over the unresolved content. Based on this view, we introduce *Late Mask Entropy* (LME), which summarizes masked-region uncertainty near the end of generation, and *Token Informativeness* (TI), which adaptively identifies sparse commitment events that most strongly reshape uncertainty over the remaining masked tokens. Across three DLMs, four QA benchmarks, and multiple decoding settings, our approach provides competitive hallucination detection using a single generation, without training an auxiliary detector or modeling the full denoising trajectory.

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 through next-token prediction, DLMs generate through iterative bidirectional denoising, repeatedly predicting and remasking tokens across the sequence. This generation process enables flexible token commitment orders and parallel decoding, while recent DLMs have demonstrated competitive performance across a range of language tasks (Bie et al., 2025).

Despite these advances, DLMs remain susceptible to hallucination, producing linguistically plausible yet factually incorrect content (Huang et al., 2025). Existing hallucination detection methods for AR-LLMs typically operate on output-level or single-step generation signals [CITE]. These approaches are designed around left-to-right generation and therefore do not directly capture the multi-step denoising dynamics of DLMs. Recent DLM-specific detectors address this limitation by modeling the denoising trajectory, for example through selected sub-traces (Chang et al., 2025) or temporal token interactions (Hemmat et al., 2026). Although effective, these approaches rely on supervised detector training with labeled data, as well as specialized trajectory modeling. This raises a natural question: can useful hallucination evidence be recovered directly from the denoising trajectory without learning a specialized detector?

In this work, we first observe an early factual commitment phenomenon in DLMs, where factual responses tend to reveal answer-defining information earlier than hallucinated responses. This suggests that factuality is not uniformly expressed throughout the denoising trajectory, but can emerge through *a small number of consequential token commitments*. To capture this behavior, we

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

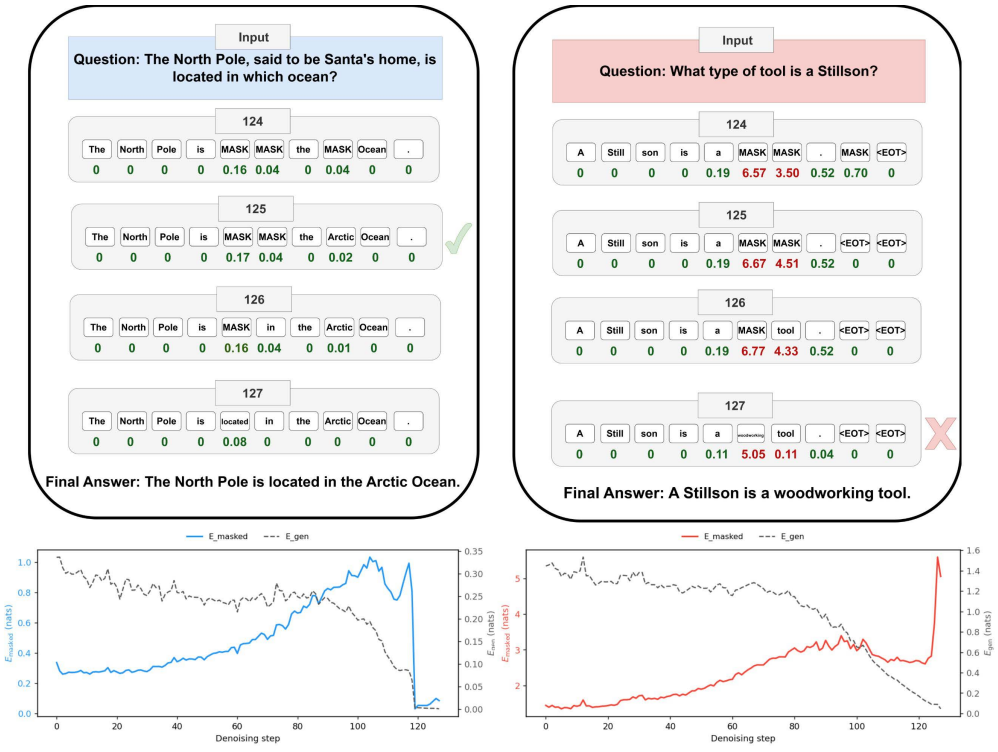


Figure 1: **Illustration of token-level trajectory behavior in factual versus hallucinated generation.** The top plots show the entropy of masked tokens across remasking steps. The bottom plots show the earliest step at which the partially generated response becomes identifiable as factual or hallucinated. 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.

separate each intermediate state into two complementary regions, the masked region, which represents unresolved content, and the generated region, which contains already committed context. We find that masked-region entropy provides a direct measure of uncertainty over the remaining response. Under confidence-based remasking, uncertain tokens are preferentially deferred toward the end of generation, making late-stage masked entropy a strong hallucination signal.

Building on this observation, we first introduce *Late Mask Entropy* (LME), a simple and efficient detector that aggregates masked-region entropy over the final few remasking steps. While LME provides a strong signal from a fixed late-stage window, informative commitments may occur at different points in the denoising trajectory. To complement this fixed-window detector, we additionally introduce *Token Informativeness* (TI), which adaptively identifies consequential commitment events based on how strongly each commitment changes uncertainty over the remaining masked tokens. By selecting a sparse set of high-impact commitments according to their uncertainty dynamics, TI captures informative events regardless of when they occur in the trajectory.

Experiments with LLaDA (Bie et al., 2025; 2026), Dream (Ye et al., 2025), and DiffusionGemma (DiffusionGemmaTeam et al., 2026) across four QA datasets and two decoding schedules demonstrate competitive response-level hallucination detection. LME achieves up to 88.26 AUROC on LLaDA, while TI consistently outperforms LME across all six DiffusionGemma settings, with gains of up to **+26.22** points. Together, these results show that useful hallucination evidence can be recovered from sparse masked-region commitment dynamics using a single generation, without auxiliary verifiers, repeated sampling, or dense trajectory modeling.

Our key contributions are summarized as follows:

- We identify early factual commitment in LLaDA, showing that factual responses tend to become identifiable earlier than hallucinated responses. This reveals that factuality signals are distributed non-uniformly across the denoising trajectory and motivates region-specific uncertainty analysis.
- Building on this observation, we develop two lightweight hallucination detectors based on commitment-level uncertainty. *Late Mask Entropy* (LME) aggregates masked-region uncertainty over the final remasking steps, while *Token Informativeness* (TI) adaptively selects a sparse set of high-impact commitments according to an interpretable entropy-change criterion.
- We demonstrate competitive hallucination detection across three DLMs and three QA datasets, with improved robustness across varying commitment timing and without training an auxiliary detector.

2 Related Work

Diffusion Language 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 these advances, DLMs remain susceptible to hallucination, motivating methods that better characterize and detect factual errors throughout the denoising process.

Hallucination Detection. has been widely studied for autoregressive LLMs using prompt-based self-evaluation, token-level 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)[PLEASE CITE SEPARATELY]. These methods are largely developed around left-to-right generation, where each token is produced once conditioned on a fixed prefix. However, DLMs generate without a fixed left-to-right order through iterative bidirectional denoising, where tokens may be repeatedly predicted, remasked, and committed throughout the sequence. Recent DLM-specific detectors therefore model the denoising 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. Although effective, these approaches require supervised detector training and additional trajectory modeling. In contrast, we investigate whether hallucination-relevant evidence can be recovered from sparse token-level commitment signals, without learning a specialized trajectory representation. Additional related work is discussed in Appendix A.

3 Method

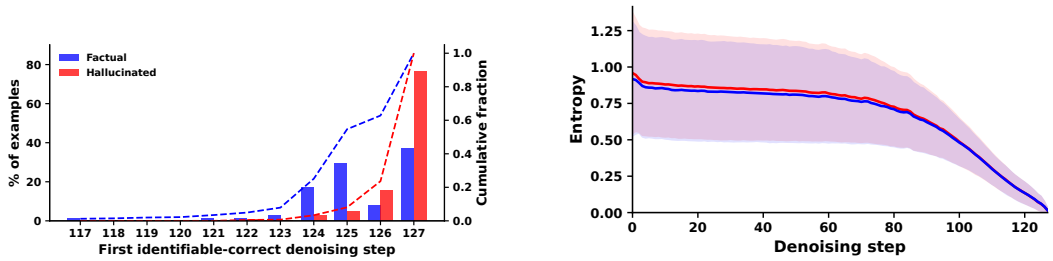
Notation. A diffusion language model begins with a fully masked response region of length L and iteratively reveals tokens over T remasking steps. We use commit to refer to the event in which a token position transitions from masked to a fixed vocabulary token.

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

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

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:

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

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

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

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.

Model	Steps	Last 1	Last 5	Last 10
LLaDA-8B-Instruct	128	80.02	**87.14**	85.88
	64	81.42	**86.73**	84.35
Dream-7B-Instruct	128	77.40	**86.16**	84.51
	64	71.71	**86.04**	83.99
DiffusionGemma-26B	128	59.26	59.21	**59.31**
	64	58.41	**58.51**	58.28

Table 1: AUROC on TriviaQA using entropy averaged over the last 1, 5, or 10 available steps. **Steps** denotes the generation schedule length.

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:

$$S_{\mathrm{LME}}^{(m)} = \frac{1}{m} \sum_{j=1}^{m} E_{\mathrm{masked}}^{(T-j)}. \tag{5}$$

In our experiments, we use $m = 5$ for LME. Table 1 shows that averaging entropy over the final five available steps achieves the best AUROC for LLaDA and Dream under both generation schedules. For DiffusionGemma, however, all three late-step windows perform similarly poorly, suggesting that the usefulness of late-step entropy depends on the model backbone and its denoising procedure. This motivates a broader question: *Can we identify informative steps throughout the denoising trajectory without restricting aggregation to a fixed late-step window?*

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.

Table 2: AUROC(%) comparison of hallucination detection methods. 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. The highest score is **bolded** and the second highest is underlined.

Model	Method	SS	TriviaQA		HotpotQA		CommonsenseQA	
			128	64	128	64	128	64
LLaDA-8B-Instruct	*Training-Free*							
	Perplexity	✗	73.18	69.74	71.03	75.15	75.85	**74.70**
	LN-Entropy	✓	61.82	66.22	59.94	70.70	65.51	66.41
	Semantic Entropy	✗	71.95	84.69	70.67	82.59	58.46	56.97
	Lexical Similarity	✗	71.52	76.28	71.90	75.99	58.38	57.02
	EigenScore	✗	62.55	78.04	64.38	81.84	58.57	57.00
	Ours-LME	✓	**87.14**	**86.73**	**86.11**	**88.26**	**79.88**	74.17
	Training-Based							
	TSV	✓	66.73	65.77	60.71	74.15	62.23	56.00
	TraceDet	✓	72.75	81.63	75.68	84.49	78.58	78.79
	Ours-TI	✓	86.44	**86.61**	**85.20**	**86.56**	**82.33**	**80.56**
	DynHD	✓	**87.63**	83.96	84.15	84.46	79.19	78.79
Dream-7B-Instruct	*Training-Free*							
	Perplexity	✗	82.87	83.95	73.03	81.22	65.06	60.31
	LN-Entropy	✓	60.75	61.94	54.12	61.99	62.09	62.73
	Semantic Entropy	✗	68.46	72.85	70.42	76.27	62.70	60.72
	Lexical Similarity	✗	69.42	67.87	71.81	73.46	63.48	61.26
	EigenScore	✗	67.67	71.36	69.79	77.51	62.00	59.54
	Ours-LME	✓	**86.16**	**86.04**	**80.36**	**84.01**	**79.72**	**70.92**
	Training-Based							
	TSV	✓	83.09	**87.03**	**80.43**	**83.72**	55.21	53.05
	TraceDet	✓	78.37	85.42	79.11	78.28	74.67	72.94
	Ours-TI	✓	**84.24**	78.69	79.87	79.65	**80.10**	**73.74**
	DynHD	✓	82.81	82.70	78.75	76.87	78.99	71.20
DiffusionGemma-26B	*Training-Free*							
	Perplexity	✗	76.82	77.93	75.89	79.20	**74.15**	**72.24**
	LN-Entropy	✓	66.26	68.59	65.60	67.43	41.38	42.95
	Semantic Entropy	✗	66.84	67.30	64.43	67.14	51.97	52.02
	Lexical Similarity	✗	69.60	72.08	66.89	67.56	52.24	52.42
	EigenScore	✗	63.25	66.41	62.78	62.77	52.26	52.43
	Ours-LME	✓	59.21	58.51	62.86	58.50	55.42	51.78
	Training-Based							
	TSV	✓	84.03	83.37	71.76	74.17	59.98	58.53
	TraceDet	✓	84.24	84.02	**83.73**	**84.10**	73.57	71.09
	Ours-TI	✓	**84.24**	**84.36**	81.98	81.03	**74.62**	**71.98**
	DynHD	✓	83.38	82.89	82.78	83.51	72.24	69.53

4 EXPERIMENTS

4.1 SETUP

Datasets. We evaluate hallucination detection on three factuality-oriented QA benchmarks and one hallucination specific dataset: TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018), CommonsenseQA (Talmor et al., 2019), and TruthfulQA (Lin et al., 2022).

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 training, we follow the same pipeline as previous works and use a training set of up to 1000 samples and evaluate on 600 held-out examples. Unless otherwise stated, all reported AUROC scores are computed on the 600 samples held-out split.

Models. We evaluate on three representative diffusion language model backbones: LLaDA-8B-Instruct (Nie et al., 2025), Dream-7B-Instruct (Ye et al., 2025), and GemmaDiffusion-26B-A4B (DiffusionGemmaTeam et al., 2026). We evaluate each model under 64-step and 128-step remasking schedules to test whether the proposed signals remain robust across generation lengths. As Gemma

Table 3: AUROC (%) of four detectors trained with 150 examples, using 128 denoising steps. Each entry is the mean ± standard deviation over four independent random training draws. The highest mean for each model and dataset is **bolded**.

Model	Method	TriviaQA	HotpotQA	CSQA	TruthfulQA
LLaDA-8B-Instruct	TSV	$61.57_{\pm 5.56}$	$60.22_{\pm 5.42}$	$61.09_{\pm 1.94}$	$63.34_{\pm 1.41}$
	TraceDet	$69.23_{\pm 1.62}$	$73.77_{\pm 1.56}$	$75.33_{\pm 1.08}$	$51.79_{\pm 1.51}$
	Ours-TI	$86.43_{\pm 0.02}$	$\mathbf{85.21}_{\pm 0.00}$	$\mathbf{82.33}_{\pm 0.00}$	$\mathbf{76.66}_{\pm 0.00}$
	DynHD	$\mathbf{87.12}_{\pm 2.67}$	$83.16_{\pm 0.80}$	$77.75_{\pm 1.49}$	$69.35_{\pm 2.86}$
Dream-7B-Instruct	TSV	$76.93_{\pm 5.08}$	$71.23_{\pm 6.48}$	$53.94_{\pm 0.79}$	$64.79_{\pm 3.91}$
	TraceDet	$76.03_{\pm 2.26}$	$75.81_{\pm 0.71}$	$68.96_{\pm 3.07}$	$66.17_{\pm 1.24}$
	Ours-TI	$\mathbf{84.24}_{\pm 0.01}$	$\mathbf{79.84}_{\pm 0.02}$	$\mathbf{80.11}_{\pm 0.01}$	$\mathbf{75.78}_{\pm 0.02}$
	DynHD	$83.30_{\pm 0.60}$	$78.90_{\pm 1.88}$	$77.49_{\pm 0.45}$	$58.87_{\pm 5.26}$
DiffusionGemma-26B	TSV	$70.21_{\pm 2.40}$	$68.39_{\pm 8.63}$	$60.21_{\pm 0.87}$	$\mathbf{72.32}_{\pm 2.74}$
	TraceDet	$81.38_{\pm 1.01}$	$79.86_{\pm 1.16}$	$65.63_{\pm 3.29}$	$68.04_{\pm 0.76}$
	Ours-TI	$\mathbf{84.24}_{\pm 0.00}$	$\mathbf{82.00}_{\pm 0.01}$	$\mathbf{74.68}_{\pm 0.14}$	$67.95_{\pm 0.01}$
	DynHD	$82.47_{\pm 0.83}$	$80.20_{\pm 1.11}$	$52.87_{\pm 9.67}$	$68.04_{\pm 0.15}$

revises a whole canvas and uses no mask, we substitute the entropy at its fixed answer positions for the masked-position entropy. We include further studies on LLaDA-2.1-mini (Bie et al., 2026) in Appendix E.

Response labeling. For each response, 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 (if available), and returns a binary judgment indicating whether the response is factually correct with respect to the reference. 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 D.

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), TraceDet (Chang et al., 2025), and DynHD (Qian et al., 2026). For evaluation, we include only baselines whose implementations were publicly available at the time of experiments.

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 compares hallucination detection methods across three datasets, three models, and two remasking lengths, with 1,000 training examples for the training-based methods. Our masked-region signals perform strongly on LLaDA and Dream: Ours-LME achieves the highest training-free AUROC in 11 of their 12 settings, outperforming the strongest prior training-free baseline by up to 14.66 percentage points. Ours-TI, achieves the highest training-based AUROC in 12 of 18 settings. Consistent with our previous observation, LME performance nevertheless depends on the backbone and is reduced to near chance on DiffusionGemma, whereas Ours-TI remains competitive.

Table 3 examines performance with only 150 training examples. Ours-TI achieves the highest mean AUROC in 10 of 12 settings. On TriviaQA, HotpotQA, and CommonsenseQA, its 128-step AUROC remains close to the corresponding 1,000-example results. Meanwhile, Ours-TI remains competitive on the more difficult TruthfulQA dataset. Additional results across training-set sizes and remasking strategies appear in Appendix B and Appendix C, respectively.

Table 4: Compute estimates and observed runtime for detector fitting routines on TriviaQA with LLaDA-8B-Instruct, using a 1,000-example training split. Generation and feature extraction are excluded.

Method	Compute (GFLOPs)	Time (s)
Ours-TI	< 0.01	0.0669
TSV	< 0.01	0.0065
TraceDet	521	10.9
DynHD	6,929	54.7

Overall, these results support the usefulness of masked-region uncertainty at selected denoising steps for hallucination detection, particularly when training data are limited. They also support our previous hypothesis that sparse uncertainty summaries can be sufficient in hallucination detection on DLMs.

Computation Cost. We report detector fitting costs in Table 4 on TriviaQA with LLaDA-8B-Instruct at 128 denoising steps, using a nominal 1,400-example training split and cached features. TI and TSV each require less than 0.01 GFLOPs of fitting arithmetic, whereas the profiler records approximately 521 GFLOPs for TraceDet and $6{,}929$ GFLOPs for DynHD. Observed runtimes are 66.9 ms, 6.5 ms, 10.9 s, and 54.7 s, respectively. These results indicate that TI's scalar normalization and validation-based weight selection incur modest fitting cost. The table details differences in counting and evaluation scope; generation and feature extraction are excluded throughout.

Case Study on LLaDA-2.1-mini As DLMs adopt more flexible token-editing mechanisms, their decoding dynamics may offer additional signals for hallucination detection. We evaluate several variants of TI on LLaDA-2.1-mini and find that incorporating Consistency can improve the reliability of the resulting detection signal. Full results and a detailed analysis are provided in Appendix E.

5 Conclusion

We studied hallucination detection in diffusion language models through denoising-time uncertainty dynamics. Our LLaDA analysis identifies *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, isolating predictive uncertainty over unresolved content. *Late Mask Entropy* aggregates this uncertainty near the end of denoising and provides effective detection in our LLaDA and Dream experiments. Furthermore, we introduce *Token Informativeness* to reduce its dependence on fixed timesteps by selecting commitment events according to the magnitude of masked-region entropy changes, retaining an interpretable selection criterion regardless of when these events occur.

Across three DLM backbones, these signals enable competitive response-level hallucination detection, with TI substantially improving over fixed late-step aggregation in settings where LME degrades. The detector requires only a single generation and lightweight scalar calibration, without auxiliary verifiers, repeated sampling, or learned detector architectures. Performance is largely insensitive to the labeled-data budget within the evaluated range, with competitive results on TriviaQA, HotpotQA, and the smaller, open-ended TruthfulQA benchmark. Our LLaDA-2.1-mini analysis further suggests that generated-region entropy can provide complementary evidence under token-editing decoding. Together, these findings support sparse, commitment-level uncertainty aggregation as a lightweight alternative to learned trajectory modeling.

Nonetheless, we identified several challenges. Detection performance is lower on TruthfulQA than on TriviaQA and HotpotQA, highlighting the difficulty of identifying hallucinations in more open-ended settings. As DLMs evolve to incorporate more flexible token-editing mechanisms and decoding policies, further work is needed to understand how these changes affect the reliability and calibration of denoising-time uncertainty signals.

AI USE DISCLOSURE

In this work, we used generative AI tools such as *Claude Opus* for implementing our methods in code, create or modify scientific figures or tables, and suggest experimental parameters. In addition, we used generative AI tools to correct grammar and refine written sentences in the paper. We have not used generative AI tools for other tasks with required disclosure. We have reviewed all AI-assisted work: LLM-generated code was manually verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We will release all code, data, and artifacts required to reproduce every number in this paper. This comprises: (i) generation scripts and the exact prompt files for all three backbones; (ii) the fixed 1000/300/300 question assignments used by every method, so that train, validation, and test membership is identical across methods and budgets; (iii) the per-prompt denoising traces from which all entropy-based signals are computed, together with the cached encoder representations used by the embedding-based baselines; (iv) the complete set of LLM judgments used as correctness labels, including the judge model identifiers, prompts, and per-example rationales for reproducibility.

REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, 2023. doi: 10.18653/v1/2023.findings-emnlp.68.

William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity and learning, 2001. URL https://arxiv.org/abs/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, Jianfeng Tan, Lanning Wei, Ji-Rong Wen, Yipeng Xing, Xiaolu Zhang, Junbo Zhao, Da Zheng, Jun Zhou, Junlin Zhou, Zhanchao Zhou, Liwang Zhu, and Yihong Zhuang. Llada2.0: Scaling up diffusion language models to 100b, 2025. URL https://arxiv.org/abs/2512.15745.

Tiwei Bie, Maosong Cao, Xiang Cao, Bingsen Chen, Fuyuan Chen, Kun Chen, Lun Du, Daozhuo Feng, Haibo Feng, Mingliang Gong, Zhuocheng Gong, Yanmei Gu, Jian Guan, Kaiyuan Guan, Hongliang He, Zenan Huang, Juyong Jiang, Zhonghui Jiang, Zhenzhong Lan, Chengxi Li, Jianguo Li, Zehuan Li, Huabin Liu, Lin Liu, Guoshan Lu, Yuan Lu, Yuxin Ma, Xingyu Mou, Zhenxuan Pan, Kaida Qiu, Yuji Ren, Jianfeng Tan, Yiding Tian, Zian Wang, Lanning Wei, Tao Wu, Yipeng Xing, Wentao Ye, Liangyu Zha, Tianze Zhang, Xiaolu Zhang, Junbo Zhao, Da Zheng, Hao Zhong, Wanli Zhong, Jun Zhou, Junlin Zhou, Liwang Zhu, Muzhi Zhu, and Yihong Zhuang. Llada2.1: Speeding up text diffusion via token editing, 2026. URL https://arxiv.org/abs/2602.08676.

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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Shenxu Chang, Junchi Yu, Weixing Wang, Yongqiang Chen, Jialin Yu, Philip Torr, and Jindong Gu. Tracedet: Hallucination detection from the decoding trace of diffusion large language models, 2025. URL https://arxiv.org/abs/2510.01274.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Zj12nzlQbz.

DiffusionGemmaTeam, Adrien Ali Taïga, James Assiene, Daniele Calandriello, Rahma Chaabouni, João Gante, Tamara von Glehn, Nate Keating, Chris Knutsen, Martin Kukla, Tianlin Liu, Ivan Lobov, Ofir Nabati, João Gabriel Oliveira, Nicolas Perez-Nieves, Nastasia Prutianova, Bobak Shahriari, Jean Tarbouriech, Pavel Tyletski, Çağlar Ünlü, Cindy Wu, Glenn Cameron, Jerome Connor, Sertan Girgin, Maarten Grootendorst,

Alon Levkovitch, Eliya Nachmani, Omar Sanseviero, Piotr Stanczyk, Quentin Berthet, Andrew Campbell, Clément Crepy, Valentin De Bortoli, Arnaud Doucet, Romuald Elie, Alexandre Galashov, Klaus Greff, Alexis Jacq, David Ruhe, Yu-Han Wu, Sebastian Flennerhag, Brendan O'Donoghue, George Scrivener, and Shantanu Thakoor. Diffusiongemma technical report, 2026. URL https://arxiv.org/abs/2608.00146.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9367–9385, 2024. doi: 10.18653/v1/2024.findings-acl.558.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Arshia Hemmat, Philip Torr, Yongqiang Chen, and Junchi Yu. Tdgnet: Hallucination detection in diffusion language models via temporal dynamic graphs, 2026. URL https://arxiv.org/abs/2602.08048.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/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. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, 2017. URL https://arxiv.org/abs/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, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022.

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

Pengxiang Li, Yefan Zhou, Dilxat Muhtar, Lu Yin, Shilin Yan, Li Shen, Soroush Vosoughi, and Shiwei Liu. Diffusion language models know the answer before decoding, 2026. URL https://arxiv.org/abs/2508.19982.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-lm improves controllable text generation, 2022. URL https://arxiv.org/abs/2205.14217.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022. URL https://arxiv.org/abs/2109.07958.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=DWkJCSxKU5.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction, 2021. URL https://arxiv.org/abs/2002.07650.

Potsawee Manakul, Adian Liusie, and Mark J. F. 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*, pages 9004–9017, 2023. doi: 10.18653/v1/2023.emnlp-main.557.

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, 2025. URL https://arxiv.org/abs/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, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd,

Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. Steer llm latents for hallucination detection, 2025. URL https://arxiv.org/abs/2503.01917.

Yanyu Qian, Yue Tan, Yixin Liu, Wang Yu, and Shirui Pan. Dynhd: Hallucination detection for diffusion large language models via denoising dynamics deviation learning, 2026. URL https://arxiv.org/abs/2603.16459.

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

Jie 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, 2023. URL https://arxiv.org/abs/2209.15558.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Haoran Sun, Lijun Yu, Bo Dai, Dale Schuurmans, and Hanjun Dai. Score-based continuous-time discrete diffusion models, 2023. URL https://arxiv.org/abs/2211.16750.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019. URL https://arxiv.org/abs/1811.00937.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

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, 2018. URL https://arxiv.org/abs/1809.09600.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models, 2025. URL https://arxiv.org/abs/2508.15487.

Appendix

CONTENTS

1 Introduction **1**

2 Related Work **3**

3 Method **3**

3.1 Early Factual Commitment 4

3.2 Predictive Uncertainty 4

3.3 Mask Entropy and Generation Entropy 5

3.4 Informativeness and Consistency 5

3.5 Hallucination Detector 6

4 Experiments **7**

4.1 Setup 7

4.2 Main Results 8

5 Conclusion **9**

A Additional Related Work **15**

B Ablation Study on Training Set Size **15**

C Remasking Strategy **15**

D LLM Labeling **16**

E Additional Analysis on LLaDA-2.1-Mini **16**

F Additional Analysis on Early Factual Commitment **18**

G Token Selection via Informativeness Score **19**

Table 5: 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

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 variability (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 Ablation Study on Training Set Size

C Remasking Strategy

In this section, we further evaluate whether our signals depend on a specific 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 5.

As shown in Table 5, LME is highly effective under low-confidence remasking, where the final masked positions are explicitly enriched for uncertain tokens. However, this signal is strongly tied to

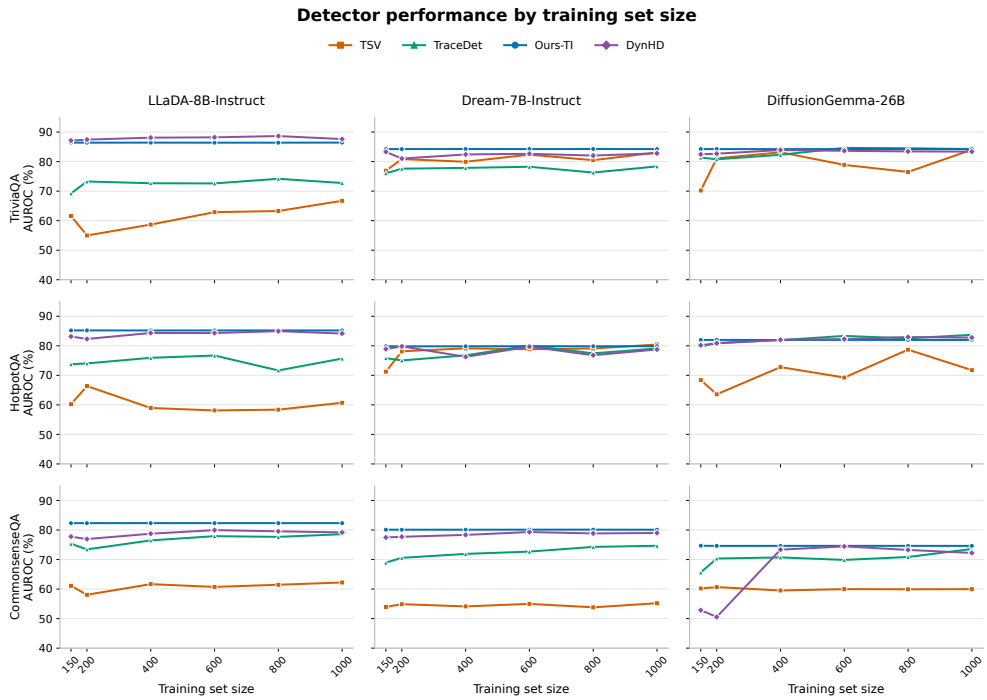


Figure 3: AUROC with of trainable methods with varying training size budget.

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.

D 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 6, 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 7.

E ADDITIONAL ANALYSIS ON LLADA-2.1-MINI

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 recent diffusion

Table 6: 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%

Table 7: 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

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 E.

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.

In this section, we provide additional analysis of LLaDA-2.1-mini. 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 Table 8, 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.

Furthermore, in Figure 4a, 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-

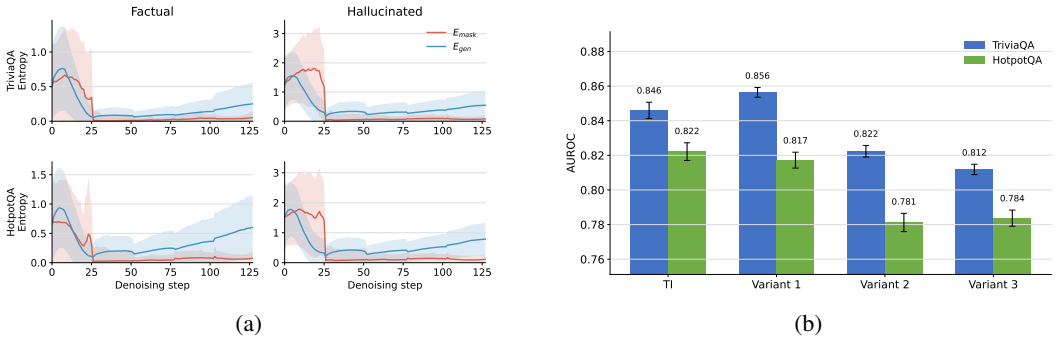
Table 8: AUROC of LLaDA-2.1-mini detector variants on 2000-questions subsets of TriviaQA and HotpotQA. Error bars are 95% confidence intervals over 20 trials.

Method	TriviaQA	HotpotQA
TI	84.60 ± 0.47	$\mathbf{82.21 \pm 0.51}$
Variant 1	$\mathbf{85.64 \pm 0.28}$	81.72 ± 0.46
Variant 2	82.24 ± 0.33	78.12 ± 0.53
Variant 3	81.19 ± 0.30	78.37 ± 0.47

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 8 and Figure 4b 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.



(a) (b)

Figure 4: 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.

F 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 9.

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.

Table 9: 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...`

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.

G 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.

	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 10: 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.

As shown in Table 10, 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.