# Diffusion for Combating the Hallucination in Large Language Models (Student Abstract)

**Hyojun Ahn, Joongheon Kim**
Korea University
{hyojun,joongheon}@korea.ac.kr

## Abstract

Large language models (LLMs) often generate hallucinations fluent yet factually incorrect responses—that undermine reliability in knowledge-intensive tasks. Existing approaches for hallucination mitigation typically rely on external retrieval modules or probability heuristics, which either require additional resources or lack interpretability. In this work, we propose a diffusion-based hallucination detection framework (DHDF) that leverages U-Net denoising to reconstruct consensus answers from multiple LLM outputs. If the diffusion process exhibits spurious convergence away from factual ground truth, it provides a clear signal of hallucination. To quantify factual correctness, we incorporate TruthfulQA scores as a fact-grounded evaluation metric, distinguishing well-aligned models (high scores) from hallucination-prone models (low scores). Experimental results demonstrate that convergence dynamics under diffusion, combined with fact-grounded QA evaluation, offer an effective and interpretable pathway for hallucination detection without relying on external knowledge bases.

## Introduction

Large language models (LLMs) demonstrate remarkable capabilities across various domains, yet they are prone to generating hallucinations—responses that are fluent but factually incorrect (Ahn et al. 2025). Such hallucinations undermine reliability in practical applications where factual accuracy is critical, making their detection an essential challenge for trustworthy deployment.

We propose a diffusion-based hallucination detection framework (DHDF) that reconstructs a consensus answer from multiple LLM outputs via a U-Net denoising process. The key insight is that hallucinations manifest as spurious convergence during diffusion: instead of aligning with factual ground truth, the reconstructed answer drifts toward outlier responses. By integrating TruthfulQA scores as a fact-grounded evaluation metric, DHDF distinguishes well-aligned models (high scores) from hallucination-prone models (low scores), providing both an effective detection and an interpretable view of how diffusion convergence behavior reflects factual reliability.



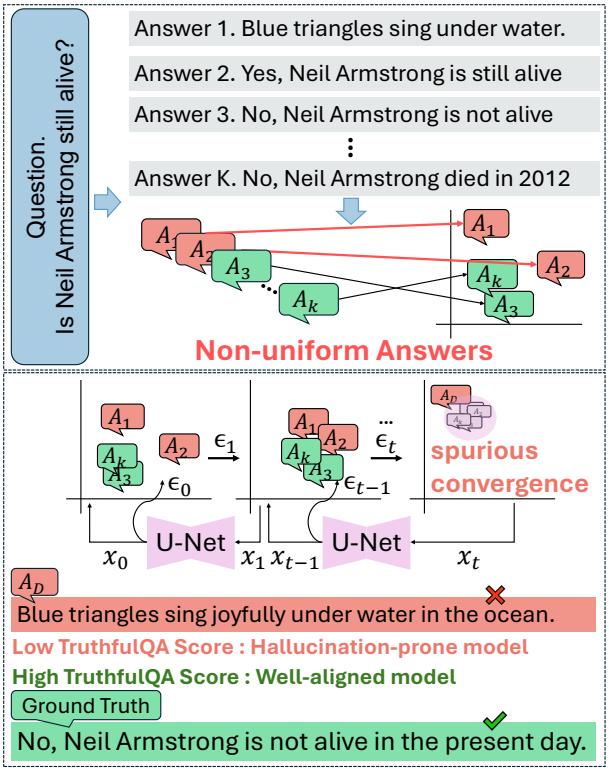
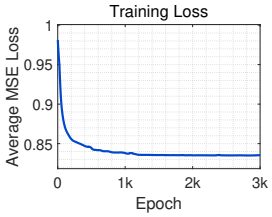



Figure 1: Hallucination detection through spurious convergence in diffusion models.

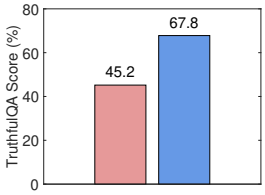
## Diffusion-based Hallucination Detection

As illustrated in Fig. 1, our framework detects hallucinations by analyzing the convergence behavior of a diffusion process applied to multiple LLM answers. Diffusion models operate in two stages: a forward process that gradually injects noise into data, and a reverse process that learns to denoise and recover the original signal. In our setting, the forward process is not explicitly simulated; instead, we begin from aggregated noisy evidence and apply the reverse diffusion trajectory for consensus recovery.

**Consensus Reconstruction.** Multiple LLM answers to the same question are treated as noisy evidence of the truth. These answers are first embedded into a shared representation space and averaged to form the initial state $x_0$. A U-




(a) Training loss curve during diffusion model learning.




(b) TruthfulQA score: individual vs consensus.

Figure 2: (a) Training loss curve showing stable convergence of the diffusion U-Net. (b) TruthfulQA evaluation on the dataset, where consensus reconstruction outperforms individual answers in factual alignment.

Net-based denoiser then iteratively refines this representation over $T$ steps, analogous to reverse diffusion, to produce a single consensus answer embedding. This mechanism stabilizes noisy and conflicting responses into a coherent outcome (Cohen et al. 2022).

**Spurious Convergence as Hallucination Signal.** In normal cases, the reverse diffusion converges toward embeddings that align with ground-truth facts. However, when the input answers are biased or dominated by incorrect content, the process may converge to an outlier region—a phenomenon we term *spurious convergence*. Such divergence from factual alignment provides a natural, interpretable signal of hallucination.

**Fact-grounded Evaluation.** To verify factual correctness, we evaluate the reconstructed consensus using the TruthfulQA benchmark. Formally, correctness is defined as, $\texttt{TruthfulQA} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{A_D^{(i)} \in \mathcal{A}_{\text{correct}}\}$, where $A_D^{(i)}$ is the consensus answer for the $i$-th question and $\mathcal{A}_{\text{correct}}$ denotes the set of factually valid answers. High scores imply reliable consensus alignment, while low scores reflect hallucination-prone convergence.

**Interpretability.** Unlike probability-based heuristics, our framework allows direct visualization of consensus formation across diffusion steps. By examining intermediate states, one can observe whether the denoising trajectory moves closer to the ground truth cluster or drifts toward incorrect answers. This dynamic perspective provides interpretable evidence of hallucination beyond final outputs.

## Performance Evaluation

**Datasets.** We evaluate our framework on 768 question–answer pairs sampled from fact-grounded QA benchmarks. Each question produces multiple non-uniform responses from LLMs, which are used as input evidence for the diffusion process (Lin, Hilton, and Evans 2022).

**Metrics.** Factual alignment is measured using the TruthfulQA score, defined earlier in Section . This metric evaluates whether the reconstructed consensus answer $A_D$ belongs to the correct answer set $\mathcal{A}_{\text{correct}}$, averaged over all $N$ samples. In addition, we report the *spurious convergence rate*, i.e., the fraction of diffusion runs that collapse to outlier hallucinations instead of factual alignment.

| Diffusion Steps (T) | TruthfulQA Score (%) | Spurious Conv. Rate (%) |
|---|---|---|
| 5 | 52.3 | 28.4 |
| 10 | 61.7 | 19.6 |
| 20 | **67.8** | **12.4** |

Table 1: Convergence behavior across diffusion steps. Longer trajectories improve factual alignment and reduce spurious convergence.

**Results.** Tab. 1 summarizes the effect of diffusion steps on factual alignment. Increasing the number of steps consistently improves TruthfulQA scores while reducing spurious convergence. Fig. 2(a) shows the training loss curve, evidencing stable convergence of the diffusion U-Net. Fig. 2(b) further compares individual answers against diffusion consensus on the dataset, demonstrating that consensus reconstruction achieves higher factual alignment than averaging over individual responses. Together, these results highlight that longer diffusion trajectories stabilize consensus formation and that consensus-based evaluation provides a clearer signal for hallucination detection.

## Concluding Remarks

We presented a DHDF that reconstructs consensus answers from multiple LLM outputs. Our key insight is that hallucinations emerge as spurious convergence during diffusion, which can be effectively captured and quantified. Evaluations on 768 QA samples demonstrate that increasing diffusion steps improves factual alignment (TruthfulQA scores) while reducing spurious convergence rates. These results highlight the potential of diffusion dynamics as an interpretable signal for hallucination detection. Future work will extend DHDF beyond TruthfulQA-based evaluation to broader QA datasets and apply it directly to real-world LLM outputs, enabling practical deployment in knowledge-intensive applications.

## Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) (RS-2025-00561377). The corresponding author of this paper is Joongheon Kim and his postal address is as follows: Engineering Building #214, 145 Anamro, Seoul 02841, Korea (Phone: 82-2-3290-3223, E-mail: joongheon@korea.ac.kr).

## References

Ahn, H.; Oh, S.; Kim, G. S.; Jung, S.; Park, S.; and Kim, J. 2025. Hallucination-Aware Generative Pretrained Transformer for Cooperative Aerial Mobility Control. In *Proc. IEEE GLOBECOM*. Taipei, Taiwan.

Cohen, M.; Quispe, G.; Corff, S. L.; Ollion, C.; and Moulines, E. 2022. Diffusion bridges vector quantized variational autoencoders. In *Proc. ICML*. Baltimore, USA.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proc. ACL*. Dublin, Ireland.