Reasoning or Rationalization? Probing the Logic of Diffusion Models

Anonymous ACL submission

Abstract

Unlike traditional autoregressive models which rely on causal reasoning to derive answers, Masked Diffusion Language Models (MDLMs) refine all sequence positions simultaneously, raising questions about the necessity of explicit reasoning steps in non-causal architectures. In this work, we investigate the dynamics of MDLM reasoning on fact verification. We observe that MDLMs typically converge on a verdict early in the generation process, treating it as a global anchor. Crucially, we find that enforcing a reasoning-first constraint via delayed verdict unmasking actively degrades performance due to refinement drift, where local noise overrides initially accurate judgments. Interventional experiments further reveal that MDLMs prioritize global sequence consistency over factual integrity, often hallucinating justifications to rationalize incorrect verdicts. Our findings suggest that for diffusion-based architectures, prolonged deliberation can be counterproductive, as it risks diluting accurate global priors with generated noise.

1 Introduction

Traditional autoregressive large language models are built on the next-token prediction objective (Vaswani et al., 2017), which necessitates a causal, sequential approach to reasoning. In this framework, Chain-of-Thought prompting (Wei et al., 2022) is often viewed as a causal prerequisite, where intermediate reasoning steps provide the necessary context for a final verdict (Lanham et al., 2023; Turpin et al., 2023). Conversely, Masked Diffusion Language Model (MDLMs) (Sahoo et al., 2024) , such as LLaDA (Nie et al., 2025), model the conditional distribution of all sequence positions simultaneously. While MDLMs provide significant flexibility in decoding order, recent evaluations (Horvitz et al., 2025) on math and coding benchmarks suggest that any-order decoding algorithms often underperform or perform similarly to standard left-to-right sampling. This observation raises questions about how to effectively utilize the bidirectional capabilities of MDLMs for complex reasoning tasks.

In this paper, we investigate the impact of iterative refinement on the performance and consistency of MDLM reasoning traces. We evaluate LLaDA-8B (Nie et al., 2025) on the AVeriTeC dataset (Schlichtkrull et al., 2023), a standard for real-world fact verification focusing on diverse claims that are less likely to be represented in pretraining data than Wikipedia-based sources. We begin by investigating a fundamental difference between causal and non-causal architectures: the sensitivity to output ordering. In autoregressive models, generating the justification before the verdict is critical, as it allows the model to condition its final prediction on the generated reasoning traces. Reversing this order often forces the model to commit to a verdict prematurely, degrading performance (Pelrine et al., 2023). We investigate whether MDLMs, with their bidirectional attention, are susceptible to this same limitation. Our results show that LLaDA-8B is remarkably robust: it achieves high accuracy regardless of whether the justification precedes or follows the verdict, significantly outperforming the standard LLaMA 3.1 (8B) (Llama, 2024) baseline and matching the performance of Qwen3-8B (Yang et al., 2025). Crucially, however, our analysis of the output ordering reveals that even when explicitly prompted to generate justifications first, LLaDA consistently predicts the verdict within the first few diffusion steps.

This observation raises a critical question: does the subsequent refinement process improve the decision, or merely rationalize it? To investigate the relationship between deliberation length and accuracy, we constrain the decoding process from predicting a verdict until 75% of the reasoning block was complete. We observe an interesting phenomenon: this forced deliberation causes sig-

nificant accuracy decay, dropping from 90.5% to 80.7%. By monitoring the evolution of the verdict token, we find that the model often identifies the correct verdict early in the diffusion process, but subsequently flips to an incorrect prediction as it fills in the justification template. In these instances, the accumulation of lower-information tokens appears to introduce local noise, causing the model to override its initial, accurate global assessment in favor of consistency with a flawed reasoning trace.

To further probe this behavior, we employ a two-way interventional test utilizing the MDLM's ability to sample from the posterior of reasoning traces given a verdict. First, we force the model to justify "incorrect" verdicts. We find that the model maintains "logical integrity" in 44% of cases—generating a justification that still points to the correct truth despite the forced label—while rationalizing the incorrect verdict through hallucinations or logic errors in the remainder. Second, we test the model's reliance on its own generated traces by forcing it to predict the verdict from the corrupted justification traces. In these scenarios, LLaDA-8B's accuracy drops to 57.3%, significantly lower than its performance when infilling from high-quality ground-truth justifications (97.1%). These findings suggest that in diffusion-based reasoning, justifications often serve as post-hoc refinements to satisfy global sequence consistency rather than as a reliable causal foundation for the final verdict.

2 Methodology

2.1 Task and Dataset

The basis of this work is to explore the utility of MDLMs for fact verification and producing faithful justifications. For this reason, we evaluate our hypotheses on the AVeriTeC dataset (Schlichtkrull et al., 2023). AVeriTeC is a benchmark dataset of factual claims and their verdicts, along with evidence to support the verdict (in the form of question-answer pairs) and human-annotated justifications based on the evidence. Unlike Wikipedia-centric datasets (e.g., FEVER), AVeriTeC utilizes diverse claims sourced from the web, requiring models to synthesize complex, often conflicting evidence. In this analysis, we focus on the core task of fact verification: given a claim and its evidence, the model must provide a verdict (i.e., Supported or Refuted) and justification. This task is particularly suitable for MDLM analysis as it allows us to observe the tension between the model's global commitment to a verdict and the justification's local adherence to specific facts.

2.2 Model Selection

Our study centers on LLaDA-8B, the first MDLM scaled to a level comparable to modern pretrained autoregressive LLMs. Unlike autoregressive LLMs that utilize causal masking to predict the next token, LLaDA utilizes a non-causal architecture where every token in the sequence can attend to every other token. During the decoding process, the model starts from a sequence of entirely masked tokens and iteratively unmasks them by sampling from the conditional distribution $P(x_i|x_{\text{unmasked}})$. This bidirectional attention mechanism allows the model to refine the global context of the sequence (such as the final verdict) concurrently with the local details of the justification, treating the entire reasoning trace as a joint distribution. In our application of LLaDA, we unmask a single token at a time as recommended for best performance by previous studies (Horvitz et al., 2025; Nie et al., 2025).

2.3 Reasoning-as-Infilling Framework

Horvitz et al. (2025) coined the term "reasoning-as-infilling", referring to the capability of MDLMs to fill in the masked values of a template, as they are not constrained to causal (left-to-right) generation. We adopt their terminology for our work; however, the implementation of our work differs slightly from theirs. While their approach proposes separating the reasoning tokens from answer tokens via an answer delimiter, we template the output in JSON format: `{"justification": "[MASK]`$_{1\ldots L}$`", "verdict": "[MASK]"}`, where L is the specified justification length.

This formulation enables us to treat the verdict as a specific block within the sequence, allowing for exact sampling from the model's posterior over justifications conditioned on a specific answer. This approach also allows us to easily parse the outputs in a structured format, ideal for our use-case where we need to extract both the justifications and the verdicts. In theory, this can generalize to any number of key-value pairs in JSON format. We utilize a fixed reasoning budget of 48 tokens to maintain a consistent refinement window. We chose this empirically as making the justifications longer would very often result in the model leaking its verdict in the justification, e.g., by saying "this claim is false because ...", rather than providing a synthesis of the facts and approaching a final verdict.

Model	Output Order	Acc (%)
LLaMA 3.1 (8B)	$V \rightarrow J$	69.7
Qwen3-8B	$V \rightarrow J$	**89.5**
LLaDA-8B	$V \rightarrow J$	88.6
LLaMA 3.1 (8B)	$J \rightarrow V$	72.5
Qwen3-8B	$J \rightarrow V$	88.0
LLaDA-8B	$J \rightarrow V$	**90.5**

Table 1: Performance of baseline models on AVeriTeC development set. Order indicates the order the justification (J) and verdict (V) are generated in.

Model	Acc (%)
LLaDA-8B	90.5
LLaDA-8B w/ 75% justification	80.7

Table 2: Performance comparison of LLaDA being constrained to fill 75% of the justification before predicting a verdict.

3 Experiments

We evaluate our experiments on the AVeriTeC development set, focusing on binary classification (Supported vs. Refuted). Due to tokenization constraints where the label "Refuted" spans multiple tokens in LLaDA's vocabulary, we map the target labels to the single-token proxies "True" (Supported) and "False" (Refuted) to facilitate single-mask infilling. In all experiments, evidence is provided as question-answer pairs in the prompt.

3.1 Baseline Models

Our core experiments are done using LLaDa-8B (`GSAI-ML/LLaDA-8B-Instruct`). Since previous work (Nie et al., 2025) demonstrates that LLaDA-8B achieves performance parity with LLaMA 3 (8B), we select the updated LLaMA 3.1 8B (`meta-llama/Llama-3.1-8B-Instruct`) as our standard autoregressive baseline. To contrast the causal reasoning of traditional autoregressive LLMs with the non-causal refinement of MDLMs, we use Qwen3-8B (`Qwen/Qwen3-8B`) with reasoning enabled as our reasoning baseline.

3.2 Justification Ordering

It is well-established that generating reasoning traces prior to answers (e.g., Chain-of-Thought) benefits autoregressive LLM performance. While previous work (Pelrine et al., 2023) has extended these findings to fact verification, the implications for MDLMs remain under-explored. In this experiment, we question whether this explicit ordering provides similar value for MDLMs. To do this, we prompt each LLM to predict the verdict (V) and justification (J) in a particular order. The results of this experiment are shown in Table 1.

The best performance came from LLaDA-8B predicting the justification first, and in general we found that predicting justifications first resulted in better performance, except with Qwen3. We suspect this is due to Qwen3's ability to determine the verdict before answering due to its pre-reasoning mechanism, which also explains the smaller performance difference between $J \rightarrow V$ and $V \rightarrow J$ compared to the other two models.

Interestingly, we found that despite the LLM prompts explicitly saying to predict the verdict first or justification first, LLaDA would always predict the verdict in the first few diffusion steps before subsequently filling out the justification. This suggests that the verdict is sufficiently clear to the model even before it begins to write the justification, or at least the model is more confident in its prediction of the verdict than any token within the justification.

3.3 Delayed Verdict Unmasking

To investigate the relationship between deliberation length (number of steps before predicting a verdict) and accuracy, we implemented a delayed unmasking mechanism. In this setup, we constrain the model so that the verdict token cannot be unmasked until at least 75% of the justification (36 of 48 tokens) has been resolved. The results of this experiment are shown in Table 2.

This intervention revealed a notable performance trade-off. Forcing the model to wait for a majority of the justification to be unmasked led to a drop in accuracy from 90.5% to 80.7%. Our analysis of "Right $\rightarrow$ Wrong" flips indicates that the model frequently arrives at the correct verdict early in the unmasking process, but subsequently changes its prediction when forced to continue constructing the justification. These findings indicate that when the unmasking schedule is artificially constrained, extended refinement can become counter-productive. In our delayed-verdict setting, the accumulation of lower-information tokens appears to introduce refinement drift, where local inconsistencies generated during the justification process conflict with and ultimately override the model's initially correct

global assessment.

3.4 Interventional Faithfulness Analysis

To probe the causal link between justification and verdict, we performed two experiments.

Post-hoc Justification (Integrity Test). First, to test whether LLaDA will justify the verdict post-hoc, we hard-code an incorrect verdict in the output and infill the justification. We then used an LLM-based judge to categorize the reasoning traces (details in Appendix A.2). We found that in about 44% of cases, the model exhibited logical integrity, where the generated justification remained faithful to the evidence and contradicted the forced incorrect verdict. In the remaining cases, the model bent its logic to satisfy global consistency, either through logical errors (37%) or factual hallucinations (13%).

Logical Reliance (Reliance Test). Second, to test whether LLaDA truly relies on its own justifications, we hard-code the justifications using the corrupted justifications from the Integrity Test experiment. The model is then forced to infill the verdict. To prevent any verdict leakage, we masked any justification tokens for words or phrases like *true*, *false*, *supported*, *unsupported*, etc. A full list can be found in Appendix A.3. When given the corrupted justifications, accuracy dropped to 57.3% largely due to the model failing to predict supported claims correctly (17.2% accuracy) compared to refuted claims where performance was decent (73.4% accuracy). In contrast, when the model was provided with high-quality ground-truth justifications from the AVeriTeC dataset, accuracy reached 97.1%.

This disparity highlights that the model's final verdict is strongly causally dependent on the quality of the reasoning trace. When the reasoning is sound (i.e., with ground-truth justifications), the model draws the correct conclusion; when the reasoning is flawed (i.e., with corrupted justifications), the model is misled. This suggests that the "Right $\rightarrow$ Wrong" flips observed in the delayed unmasking experiment are not due to a disconnect between reasoning and answer, but rather the opposite: the model effectively "confuses itself" by generating noisy justification tokens that exert a negative causal influence, overriding its initially correct global assessment.

4 Conclusion

In this work, we investigated the utility of Masked Diffusion Language Models (MDLMs) for fact verification, with particular attention to the interplay between verdict prediction and justification generation. Our findings reveal a nuanced picture of how MDLMs handle fact verification. While LLaDA-8B achieves competitive performance when allowed to unmask freely, the model consistently prioritizes verdict prediction in early diffusion steps—even when explicitly instructed to generate justifications first. This behavior suggests that MDLMs develop strong global commitments early in the generation process, treating the verdict as a high-confidence anchor around which local details are subsequently refined.

The delayed verdict unmasking experiment identifies a critical vulnerability in this reasoning paradigm. When forced to commit to extensive justifications before settling on a verdict, model performance degrades substantially (to 80.7% accuracy), with many initially correct predictions flipping to incorrect ones. This degradation is not due to a disconnect between reasoning and answer, but rather the opposite: our interventional faithfulness analyses demonstrate that LLaDA's verdicts are strongly causally dependent on the quality of its justifications. When forced to justify incorrect verdicts, LLaDA often sacrifices factual accuracy in the justifications to ensure global consistency, maintaining logical integrity in only 44% of claims. Subsequently, when provided with these corrupted justifications, the model's accuracy plummets to 57.3% while ground-truth justifications achieve 97.1%. These results demonstrate that MDLMs' create causal dependencies between reasoning traces and answers.

Ultimately, these observations suggest that for diffusion-based architectures, maximizing inference compute via extended reasoning steps does not strictly correlate with better outcomes. Instead, there appears to be a point of "enough thinking"—where the global assessment is established but has not yet been degraded by local refinement noise. Future work should focus on dynamic inference strategies, such as early-exit mechanisms (Horvitz et al., 2025), to leverage this early global convergence while mitigating the risks of refinement drift.

Limitations

First, we acknowledge the scope of the reasoning task employed. While AVeriTeC involves real-world claims that are more complex than Wikipedia-based benchmarks like FEVER, the inclusion of gold-standard evidence in the prompt reduces the complexity. It remains an open question whether the phenomenon of refinement drift persists in domains requiring strict, multi-step deduction, such as advanced mathematics (e.g., AIME). In those settings, intermediate steps are strict logical prerequisites for the solution, whereas in fact verification, the justification often serves as a linguistic rationalization of an intuitive verdict.

Second, our analysis is constrained by the current availability of high-performing MDLMs, limiting our experiments to the LLaDA-8B architecture. It is possible that the susceptibility to local noise and refinement drift is a function of model scale; larger models with deeper capacity for global attention might maintain coherence over longer reasoning traces without succumbing to the self-confusion observed here. Furthermore, these phenomena may also be a result of the pretraining data or optimization algorithm of LLaDA.

Finally, our analysis is only applied to masked diffusion language models, namely LLaDA. While the decoding process is similar, it's unclear whether these findings can be applied to other discrete diffusion models like SEDD (Lou et al., 2024) or continuous diffusion models (Li et al., 2022).

References

Zachary Horvitz, Raghav Singhal, Hao Zou, Carles Domingo-Enrich, Zhou Yu, Rajesh Ranganath, and Kathleen McKeown. 2025. No compute left behind: Rethinking reasoning and sampling with masked diffusion models. *Preprint*, arXiv:2510.19990.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *Preprint*, arXiv:2307.13702.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022. Diffusion-lm improves controllable text generation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Team Llama. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. Discrete diffusion modeling by estimating the ratios of the data distribution. *Preprint*, arXiv:2310.16834.

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.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6399–6429, Singapore. Association for Computational Linguistics.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. *Preprint*, arXiv:2406.07524.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *Preprint*, arXiv:2305.13117.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Preprint*, arXiv:2305.04388.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Appendix

A.1 Reproducibility

All of our experimental results are based on a single run. We found no variation in outputs across runs with LLaDA, so we only used a single run for our experiments. For LLaMA 3.1 and Qwen3, we also used a single run as their performance is not a significant factor of this work. We used the default generation parameters for both LLaMA 3.1 and Qwen3. Github Copilot was used as an autocomplete tool in some of the code for this work.

A.2 Reasoning Trace Categories

After a manual review of about 50 justifications of incorrect verdicts, we landed on four main categories as shown below:

- Logical error: The justification contains reasoning errors that lead to an incorrect conclusion.
- Cherrypicking details: The justification selectively uses evidence that supports the incorrect verdict while ignoring evidence that contradicts it.
- Verdict-justification mismatch: The justification does not logically support the given verdict (i.e., the justification suggests a different verdict).
- Factual hallucination to support wrong verdict: The justification includes evidence which is not explicitly given that are used to support the incorrect verdict.
- Other (describe in a few words): The justification contains an error that does not fit into the above categories.

A.3 Masked Justification Tokens

The full list of words and phrases masked during the logical reliance experiment are as follows: *true*, *false*, *True*, *False*, *TRUE*, *FALSE*, *no evidence*, *unsupported*, *refuted*, *supported*, *support*, *refute*, *inconsistent*, *consistent*, *accurately*, *inaccurately*, *incorrectly*, *correctly*, *impossible to verify*, *impossible to determine*.