推理还是合理化？理由在用于事实核查的掩码扩散模型中的作用

Jacob Devasier德克萨斯大学阿灵顿分校 jacob.devasier@uta.edu

arXiv:2603.01190v1 [cs.CL] 1 Mar 2026

摘要

不同于按顺序生成词元，并受益于思维链等“先推理、后作答”策略的自回归模型，掩码扩散语言模型（MDLM）会同时精炼序列中的所有位置，这引发了一个问题：这类模型如何处理需要给出有理据判定的任务？在本研究中，我们探究 MDLM 在事实核查任务中的推理动态，考察理由究竟发挥了真正推理的作用，还是仅仅构成事后合理化。我们发现，MDLM 通常会在扩散过程的早期就确定判定，并将其视为全局锚点，在理由完成之前先行解析。关键的是，通过延迟揭示判定来强制施加推理优先约束，会主动降低模型表现：准确率从 86.2% 降至 71.9%，因为不断累积的理由词元会引入不一致信息，推翻原本正确的预测。干预实验表明，在 56% 的案例中，模型会为被强制给出的错误判定寻找合理化解释；判定结果还与理由质量存在很强的因果依赖（理由受损时准确率为 57.3%，而使用真实标注理由时则为 97.1%）。这种因果依赖解释了强制审慎推理为何会导致性能下降：模型生成噪声论证词元时，会以这些词元为条件逐步作答，从而推翻最初正确的判断。我们的研究发现，对于使用 MDLM 进行事实核查的任务，延长审议可能适得其反，因为理由生成过程中引入的噪声可能会稀释早期准确的预测。

1 引言

传统的自回归大语言模型基于下一词元预测目标（Vaswani 等人，2017），因此必须采用因果式的顺序生成文本。在这一框架下，思维链提示（Wei 等人，2022年）通常被视为复杂推理的因果前提，因为中间步骤为得出最终答案提供了必要的上下文（Lanham 等人，2023年；Turpin 等人，2023年）。与之相反，掩码扩散语言模型（MDLM）（Sahoo 等人，2024年），例如 LLaDA（Nie 等人，2025年），对所有序列位置的条件分布进行同步建模。尽管掩码扩散语言模型在解码顺序上具有很大的灵活性，但近期针对数学与编程基准测试的评测（Horvitz 等人，2025年）表明，任意顺序解码算法的表现往往不如或与标准的从左到右采样相近。这一发现引发了一个问题：如何有效利用掩码扩散语言模型的双向能力，尤其是在需要给出有依据决策的任务中。

本文探讨迭代精炼如何影响掩码扩散语言模型生成的推理轨迹在事实核查中的性能与一致性。我们在 AVeriTeC 数据集（Schlichtkrull 等人，2023年）上评估 LLaDA-8B（Nie 等人，2025年）。该基准数据集用于真实世界的主张核查，包含多种主张；与基于维基百科的来源相比，这些主张较不可能出现在预训练数据中。事实核查为研究掩码扩散语言模型的行为提供了一个尤为有趣的试验平台：与数学推理不同，数学推理中的中间步骤是严格的逻辑前提，而事实核查中的论证可能代表真正的证据综合，也可能是对直觉的事后合理化。

我们首先考察因果架构与非因果架构之间的一个根本差异：对输出顺序的敏感性。在自回归模型中，先生成理由再给出判定至关重要，因为这使模型能够根据生成的推理轨迹来确定最终预测。颠倒这一顺序往往会迫使模型过早作出判定，从而降低性能（Pelrine等人，2023年）。我们考察具有双向注意力机制的掩码扩散语言模型是否也会受到这一局限的影响。我们的

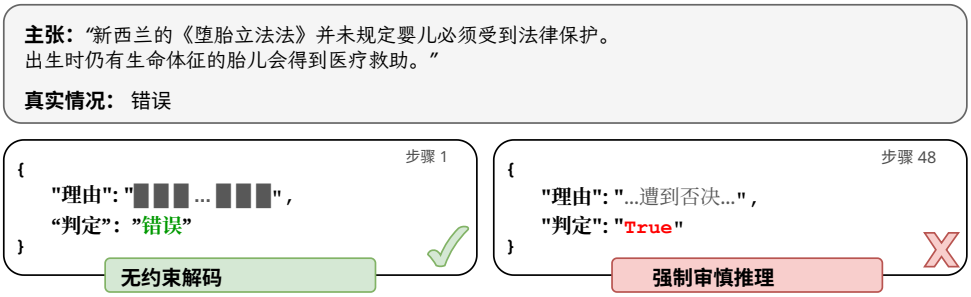


图1：细化漂移的一个示例。当模型被迫在预测判定之前生成大部分理由（无论对判定的置信度如何）时，最终会得出错误结论。

结果表明，LLaDA-8B表现出显著的稳健性：无论理由在判定之前还是之后生成，它都能取得较高的准确率，表现优于标准的LLaMA 3.1（8B）（Llama，2024年）基线模型，并达到Qwen3-8B（Yang等人，2025年）的水平。然而，关键的是，我们的分析发现，即使明确提示LLaDA先生成理由，它仍会在最初几个扩散步骤内持续预测判定。

这一观察引出了一个关键问题：后续的细化过程究竟改善了判定结果，还是仅仅生成支持性的理由？为探究这一点，我们通过阻止模型在生成指定比例的理由之前预测判定结果，来约束解码过程。事实证明，这种强制审慎推理适得其反，并未提高准确率：随着审慎推理阈值从 0% 增加到 90%，准确率持续从 86.2 % 下降至 71.9%。通过监测裁决词元在扩散步骤中的演变，我们发现，模型通常会在过程早期识别出正确的判定结果，但随着理由词元不断累积，随后又反转其预测。我们将这一现象称为 细化漂移：在生成理由的过程中逐步引入局部不一致性，这些不一致性与模型最初正确的整体评估相冲突，并最终将其覆盖。

为进一步探究理由与判定结果之间的因果关系，我们采用双向干预测试，利用掩码扩散语言模型能够以任意词元子集为条件的特性。首先，我们强制模型为错误判定生成理由。结果发现，模型仅在 44% 的情况下保持了逻辑完整性——生成的理由与强制给出的判定相矛盾，并支持正确答案；而在其余 56% 的情况下，模型会通过逻辑错误（37%）、事实性幻觉（13%）或其他失败（6%），为错误的判定进行合理化。其次，我们通过强制模型以这些被篡改的理由为条件来预测判定结果，从而检验模型对自身生成轨迹的依赖。在这种设置下，准确率骤降至 57.3%，主要原因是模型在支持的主张上几乎完全失效（准确率为 17.2%）。相比之下，若向模型提供 AVeriTeC 数据集中的高质量真实标注理由，准确率则达到 97.1%。综合来看，这些结果表明，LLaDA 的判定结果与理由质量存在很强的因果依赖，而这种依赖正是细化漂移的根源机制：随着模型生成含噪的理由词元，它会以这些词元为条件，逐渐推翻自己最初正确的评估。

2 方法论

我们在 AVeriTeC 数据集（Schlichtkrull 等人，2023年）上进行评估。该数据集包含现实世界中的事实性主张，以及与之配对的判定结果、支持性证据（以问答对的形式呈现）和人工标注的理由。不同于以维基百科为中心的数据集（如 FEVER），AVeriTeC 从多样化的网络来源中提取主张，因此要求模型综合复杂且常常相互冲突的证据。我们聚焦于核心事实核查任务：给定一项主张及其证据，模型必须同时生成判定结果（支持或驳斥）和理由。这一设定尤其适合研究掩码扩散语言模型的行为，因为它揭示了模型对判定结果的整体承诺与理由对具体事实的局部遵循之间的张力。

2.1 模型选择

我们的研究聚焦于 LLaDA-8B（Nie 等人，2025年），这是首个扩展至可与现代预训练自回归大语言模型相媲美规模的掩码扩散语言模型（MDLM），并且目前是该类别中性能最强的模型。我们还

目前是该类别中性能最强的模型。与采用因果掩码按顺序预测词元的自回归大语言模型不同，LLaDA 使用非因果架构，其中每个词元都能关注其他所有词元。这种双向注意力使模型能够同时细化全局上下文（例如判定）与局部细节（例如理由），并将整个输出视为一个联合分布。遵循先前的建议（Horvitz 等人，2025年；Nie 等人，2025年），我们在解码过程中每次仅揭示一个词元。

2.2 输出格式与生成长度 th

我们要求模型以 JSON 格式返回预测结果，以便解析判定和理由字段。我们将生成长度固定为 64 个词元；扣除 JSON 结构和裁决词元后，约有 48 个词元可用于理由。我们根据初步实验选择了这一长度：实验表明，输出较长时，理由中经常会出现判定泄漏（例如“该说法是错误的，因为……”），这会干扰我们对判定在扩散过程中的变化情况进行分析。

模型	输出顺序	准确率（%）
LLaMA 3.1（8B）	$V \to J$	69.7
LLaDA-8B	$V \to J$	86.2
Qwen3-8B	$V \to J$	89.5
LLaMA 3.1（8B）	$J \to V$	72.5
Qwen3-8B	$J \to V$	88.0
LLaDA-8B	$J \to V$	89.0

表1：基线模型在 AVeriTeC 开发集上的性能。顺序表示理由（J）和判定（V）的生成顺序。

模型	审议（%）	准确率（%）
LLaDA-8B	0	86.2
LLaDA-8B	25	86.6
LLaDA-8B	50	79.2
LLaDA-8B	75	73.8
LLaDA-8B	90	71.9

表 2：强制审慎推理对 LLaDA 性能的影响。审议百分比根据原始的 64 个推理词元跨度计算。

3 实验

我们在 AVeriTeC 开发集（$N = 500$）上进行评估，并将任务设定为二元分类：支持与驳斥。由于“Refuted”在 LLaDA 词汇表中由多个词元组成，我们将每个标签映射为单个词元代理：“True”对应支持，“False”对应驳斥。随后以问答对的形式提供证据。

3.1 基线模型

我们的主要模型是 LLaDA-8B[1]。我们将其与两个基线模型进行比较：LLaMA 3.1 8B[2], , 一种标准的自回归模型；以及启用了扩展思考模式的 Qwen3-8B[3]，作为推理增强基线。这些模型使我们能够将 LLaDA 的任意顺序生成与标准自回归解码和推理增强的自回归解码进行比较。

3.2 理由排序

在答案前生成推理轨迹（例如，思维链提示）有利于自回归模型——

有大语言模型，并已证明能够提升事实核查表现（Pelrine等人，2023年）。这种顺序优势是否也适用于无需按从左到右的顺序生成词元的掩码扩散语言模型，仍是一个悬而未决的问题。为探究这一点，我们提示每个模型按两种可能的顺序生成判定（V）和理由（J）（表1）。

三种模型对输出顺序的敏感性都很低，不过 LLaMA 3.1 的整体表现明显逊于 Qwen3-8B 和 LLaDA-8B。考察 LLaDA 的词元轨迹，有助于理解顺序为何影响甚微：无论提示的顺序如何（$J \to V$或$V \to J$），模型都会在最初几个扩散步骤内确定判定，远早于理由生成完毕。换言之，LLaDA 会立即得出答案，之后再补充推理。我们认为，这种表现源于该场景下二元事实核查的复杂度相对较低，因此模型无需经过长时间审议便能确定判定。

3.3 延迟裁决去掩码

前述结果引出了一个自然的后续问题：强制模型进行更长时间的审议，是否会提高准确率？为检验这一点，我们阻止

[1]https://huggingface.co/GSAI-ML/LLaDA-8B-Instruct[2]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct[3]https://huggingface.co/Qwen/Qwen3-8B

裁决词元被取消掩码，直到指定比例的理由词元已被揭示（表2）。与预期相反，这种干预损害了模型表现，使准确率从 86.2% 降至 71.9%。有趣的是，我们发现，只有在审议超过 25% 后，模型表现才出现显著下降。

检查判定轨迹揭示了其根本原因：模型常常在扩散过程早期就得出正确判定，但随着生成更多理由词元，它会反转预测。我们将这一现象称为细化漂移。随着理由词元不断累积，它们会引入局部不一致，并逐渐与模型最初正确的整体评估相冲突，最终导致模型推翻自己先前——且更准确——的预测。我们还发现，审议比例在 0–25% 之间时的稳定性，也与输出 JSON 结构的生成相吻合；该结构通常在最初的 10–16 个词元内生成。

3.4 干预忠实性分析

为探究理由与判定之间的因果关系，我们利用掩码扩散语言模型的一项独特特性：能够以任意词元子集为条件。这使我们能够实施两种互补的干预——一种用于检验模型是否会为被强制给出的判定寻找理由，另一种用于检验模型在预测判定时是否真正依赖理由。

完整性测试。我们硬编码一个错误的判定，并让模型补全其周围的理由。我们使用 Qwen3-30B-A3B 作为评判模型（类别详见附录A.2），发现模型在 44% 的情况下保持了逻辑完整性，生成的理由与强制给定的判定相矛盾，转而支持正确答案。在其余情况下，模型会为错误判定寻找合理解释，要么出现逻辑错误（37%），要么臆造证据中不存在的事实（13%）。[4]

依赖性测试。 接下来，我们提出相反的问题：LLaDA 在预测判定时，是否真正依赖自己生成的理由？为此，我们将完整性测试中生成的被篡改的理由作为上下文硬编码，迫使模型仅补全裁决词元。为防止理由文本泄漏判定，我们屏蔽了与标签指示性词语和短语相对应的词元，例如正确、错误、支持和不支持 （完整列表见附录A.3）。

在以被篡改的理由为条件时，准确率骤降至 57.3%。这一降幅主要源于模型在支持的主张上几乎完全失效（准确率为 17.2%），而其在被反驳的主张上的表现则仅有较温和的下降（准确率为 73.4%）；我们在附录A.4中分析了这种不对称性。相比之下，当模型获得来自 AVeriTeC 数据集的高质量真实依据时，准确率达到 97.1%。

启示。 这种差异表明，模型的判定在很大程度上因果依赖于其推理轨迹的质量。重要的是，这种依赖并不与第 3.2 节观察到的判定早期趋同相矛盾——相反，它揭示了细化漂移背后的机制（第 3.3 节）。模型会在最初几个扩散步骤内确定判定，然后生成理由词元，并在后续过程中以这些词元为条件进行推理。当这些词元包含错误时，它们便会施加错误方向的影响，逐渐推翻最初正确的评估。

4 结论

在本研究中，我们考察了 LLaDA-8B 如何在 AVeriTeC 数据集上进行事实核查。我们发现，该模型取得了有竞争力的表现，并且对输出顺序具有稳健性，但即便被指示先生成理由，它也始终在最初几个扩散步骤内就确定判定。通过延迟裁决去掩码强制延长审议被证明适得其反，准确率从 86.2% 降至 71.9%，因为不断累积的理由词元覆盖了最初正确的预测。我们的干预分析揭示了其内在机制：LLaDA 的判定结果因果依赖于理由质量。当被迫为错误判定提供理由时，该模型仅在 44% 的情况下保持逻辑完整性；而以这些被篡改的理由为条件时，准确率降至 57.3%，相比之下，使用真实依据时准确率为 97.1%。这些发现表明，对于当前的掩码扩散语言模型，在受限解码下延长审议可能会用来自理由生成过程的噪声稀释早期的准确预测，不过仍需未来的工作来确定这一模式是否能推广到其他推理任务和模型规模。

[4]其余 6% 被归为“其他”。

局限性

首先，尽管 AVeriTeC 涉及的现实世界主张比 FEVER 等基于维基百科的基准更为复杂，但在提示中纳入金标准证据降低了任务复杂度。细化漂移是否会持续存在于需要严格、多步演绎的领域（例如高等数学，如 AIME）中，仍是一个悬而未决的问题；在这些领域，中间步骤是求解的逻辑前提，而非对直觉性判定的语言合理化。

其次，我们的实验仅限于 LLaDA-8B，这是目前在该规模下唯一可用的高性能掩码扩散语言模型。对精炼漂移的易感性可能取决于模型规模、预训练数据或优化过程；具有更强全局注意力能力的更大规模模型或许能够在更长的推理轨迹上保持连贯性。

第三，我们的发现可能无法推广到掩码扩散模型之外。尽管解码过程存在相似之处，但尚不清楚诸如 SEDD（Lou 等人，2024年）或连续扩散模型（Li 等人，2022年）是否会表现出相同的行为。

参考文献

Zachary Horvitz、Raghav Singhal、Hao Zou、Carles Domingo-Enrich、ZhouYu、RajeshRanganath 与 Kathleen McKeown。2025年。不让算力闲置：借助掩码扩散模型重新思考推理与采样。预印本，arXiv:2510.19990。

Tamera Lanham、Anna Chen、Ansh Radhakrishnan、Benoit Steiner、Carson Denison、Danny Hernandez、Dustin Li、Esin Durmus、Evan Hubinger、Jackson Kernion、Kamil˙e Lukoši ̄ut˙e、Karina Nguyen、Newton Cheng、Nicholas Joseph、Nicholas Schiefer、Oliver Rausch、Robin Larson、Sam McCandlish、Sandi- pan Kundu 等另外 11 位作者。2023年。衡量思维链推理的忠实度。预印本，arXiv:2307.13702。

Xiang Lisa Li、John Thickstun、Ishaan Gulrajani、Percy Liang 和 Tatsunori B. Hashimoto。2022年。Diffusion-lm 改进了可控文本生成。载于第*36*届神经信息处理系统国际会议论文集，NIPS '22，雷德胡克，纽约州，美国。Curran Associates Inc.

Llama 团队。2024年。 Llama 3 模型群。预印本，arXiv:2407.21783。

Aaron Lou、Chenlin Meng 和 Stefano Ermon。2024年。通过估计数据分布的比率进行离散扩散建模。预印本，arXiv:2310.16834。

Shen Nie、Fengqi Zhu、Zebin You、Xiaolu Zhang、JingyangOu、JunHu、Jun Zhou、YankaiLin、Ji-Rong Wen 和 Chongxuan Li。2025年。大型语言扩散模型。预印本，arXiv:2502.09992。

Kellin Pelrine、Anne Imouza、Camille Thibault、Meilina Reksoprodjo、Caleb Gupta、Joel Christoph、Jean-François Godbout 与 Reihaneh Rabbany。2023年。迈向可靠的虚假信息缓解：泛化、不确定性与 GPT-4。载于 《*2023*年自然语言处理经验方法会议论文集》，第6399–6429页，新加坡。计算语言学协会。

Subham Sekhar Sahoo、Marianne Arriola、Yair Schiff、Aaron Gokaslan、Edgar Marroquin、Justin T Chiu、Alexander Rush 与 Volodymyr Kuleshov。2024年。简单有效的掩码扩散语言模型。 预印本，arXiv:2406.07524。

Michael Schlichtkrull、Zhijiang Guo 和 Andreas Vla-chos。2023年。Averitec：一个包含来自网络证据、用于现实世界主张核查的数据集。载于第*37*届神经信息处理系统国际会议论文集，NIPS '23，美国纽约州雷德胡克。Curran Associates Inc.

Miles Turpin、Julian Michael、Ethan Perez 和 Samuel R. Bowman。2023年。语言模型并不总是说出它们所思考的内容：思维链提示中的不忠实解释。载于第*37*届神经信息处理系统国际会议论文集，NIPS '23，美国纽约州雷德胡克。Curran Associates Inc.

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。你所需要的就是注意力机制。载于第*31*届神经信息处理系统国际会议论文集，NIPS'17，第6000–6010页，美国纽约州雷德胡克。Curran Associates Inc.

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、Ed H. Chi、Quoc V. Le 和 Denny Zhou。2022年。思维链提示在大语言模型中引发推理。载于第*36*届神经信息处理系统国际会议论文集，NIPS '22，美国纽约州雷德胡克。Curran Associates Inc.

An Yang、AnfengLi、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv、Chujie Zheng、Day- i heng Liu、Fan Zhou、Fei Huang、Feng Hu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang 以及其他41位作者。2025年。Qwen3技术报告。预印本，arXiv:2505.09388。

A 附录

A.1 可复现性

所有结果均基于单次运行。LLaDA 一次解掩码一个词元时（每一步选择概率最高的词元），其解掩码过程是确定性的，

主张	证据	真相	预测	生成的解释与错误
Amy 科尼·巴雷特 was 获确认 as 最高法院大法官 于10月26日，2020年。	**问：** 艾米·科尼·巴雷特是否获确认担任最高法院大法官？ **答：** 是。	补充材料	Ref.	“……确认日期不正确。她于2020年9月26日获得确认。” 编造了错误日期；证据中没有日期。
查谟和克什米尔从……中移除联合国“未决争端”名单。	**问：** 查谟和克什米尔曾列入安全理事会议程吗？ **答：** “查谟和克什米尔从未列入议程……它是归入单独的类别。”	Ref.	补充材料	“这项说法属实，因为该争议从未列入议程……这表明它已被移除。” 将“从未列入”与“已移除”混为一谈。

表3：依赖性测试的示例预测，展示被篡改的理由如何误导模型。Supp.= 支持，Ref. =驳斥。

因此不同运行之间会产生完全相同的输出。对于 LLaMA 3.1 和 Qwen3，我们使用了采用默认生成参数的单次运行，因为这些模型仅作为基线，并非我们分析的重点。开发过程中使用 GitHub Copilot 作为自动补全工具。

A.2 推理轨迹类别

为了对模型在完整性测试中如何处理错误的强制判定结果进行分类，我们人工审阅了约50条理由，并识别出以下错误类别，随后将其用作大语言模型评审器（Qwen3-30B-A3B）的指令：

- 逻辑错误：该理由包含导致错误结论的推理错误。
- 选择性挑选细节：该理由选择性使用支持错误判定的证据，同时忽略与之矛盾的证据。
- 判定与理由不匹配：该理由在逻辑上并不支持给定的判定（即该理由暗示了不同的判定）。
- 用以支持错误判定的事实性幻觉：该理由包含并未明确给出、却被用来支持不正确判定的证据。
- 其他（用几个词描述）：该理由包含不属于上述类别的错误。

A.3 掩码理由词元

在依赖性测试期间，我们掩码了所有对应于以下标签指示性词语的词元

以及短语，以防止通过理由文本发生判定泄漏：真，假，真，假，真，假，无证据，不受支持，驳斥，支持，支持，驳斥，不一致，一致，准确地，不准确地，不正确地，正确地，无法验证，无法确定。

A.4 支持/驳斥准确率不对称

在依赖性测试中，针对支持的主张的准确率（17.2%）远低于针对驳斥的主张的准确率（73.4%）。我们将这种不对称归因于事实核查语言的语言学特性。在掩码判定泄漏词元之后，驳斥性理由仍保留“骗局”“讽刺作品”和“假新闻”等明确信号，而支持性理由则失去了大部分判别性内容——像“已确认”这样的词在缺少周围语境时会变得含糊。这种不对称还因驳斥语言固有的直白性而进一步加剧；相比之下，支持类表述通常更为含蓄（例如，“证据表明”“尽管尚未确认”）。

模型的认同率证实了这一模式：它在85.6%的情况下遵循驳斥理由，但仅在26.6%的情况下遵循支持理由。当正向信号较弱时，模型似乎会退回到一种先验——很可能是在预训练期间习得的——即接受事实核查审视的主张往往为假。

A.5 带有被篡改理由的示例预测

表3 展示了依赖性测试中被篡改的理由误导模型的示例。