# 超越自回归：扩散语言模型是模拟存内计算的需求

Yuanuo Feng[†1]、Yubing Ye[†2]、Yizhe Chen[1],Wenshuai Yao[3],Ngai Wong[2],Wenyong Zhou[*2]、Wang Kang[*1]、[1]
北京航空航天大学集成电路科学与工程学院，中国北京[2]香港大学工程学院，香港[3]北京大学集成电路学院，中国北京 [†]:
贡献相同。 [*]：通讯作者。

arXiv:2611.32223v1 [cs.AR] 24 Nov 2026

**摘要—模拟存内计算（AIMC）能够实现高能效的语言模型推理，但硬件非理想性会扰动矩阵向量乘法并降低生成质量。自回归（AR）模型尤其容易受到影响：每个词元在生成时便被提交到因果上下文中，因此扰动引起的错误会级联传播。我们研究了扩散语言模型（DLMs）这一替代范式。扩散语言模型会对掩码序列进行迭代式细化。在规模匹配的模型对和开放检查点上，与自回归模型相比，扩散语言模型在持久性权重扰动下的性能退化更为平缓。受控分析排除了跨轮次噪声平均的影响，并表明其机制在于延迟词元承诺和渐进式上下文解析。更大的去噪步数预算还能进一步缩小干净与噪声性能差距，形成扩散语言模型特有的延迟—鲁棒性调节旋钮。我们使用基于真实硅测量结果校准的噪声模型，在多种数据集和模型上开展了大规模仿真测试。此外，在一套16芯片28纳米电可擦除闪存模拟存内计算集群上进行的混合板载测量表明，扩散语言模型能够容忍1.4–3× 更大的权重噪声，使基于扩散的生成成为抗噪模拟存内计算推理的一种有前景范式。**

索引词—**模拟存内计算、扩散语言模型、自回归生成。**

## 一、引言

大语言模型（LLM） [1] 越来越多地应用于需要低延迟 [2] 和高能效生成 [3]的场景。模拟存内计算（AIMC）是适用于这类工作负载的一种有前景的底层技术，它直接在存储器中执行矩阵向量乘法（MVM），从而减少数据搬移 [4]。然而，模拟存内计算会引入器件级和电路级非理想性，包括编程可变性、电导漂移、读取噪声以及外围电路缺陷 [5], ，这些因素会扰动有效权重，进而影响每次 MVM 的输出 [6]。尽管这类扰动

这些差异在操作层面可能很小，但会在语言生成过程中累积并放大 [4], [7], [8]。

这种放大效应对于自回归（AR）生成尤其成问题。自回归模型从左向右生成词元，每个生成的词元都会立即纳入用于预测后续所有词元的因果上下文。因此，较早位置因扰动而产生的误差可能改变后续位置所观察到的上下文，并在序列剩余部分引发错误级联 [9]–[11]。因此，模拟存内计算上的自回归鲁棒性不仅取决于每次运算或每个词元的误差，还取决于局部误差是否会不可逆地写入生成轨迹。

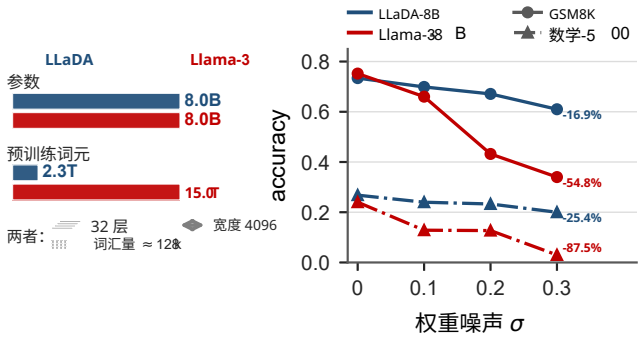


图 1. 两个 8B 规模语言模型在权重扰动下的比较。LLaDA-8B 和 Llama-3-8B 的深度（32 层）、宽度（4096）和词汇表大小（≈128k）相同。

扩散语言模型（DLM）采用不同的生成过程 [12]：它们并非严格按照从左向右的顺序确定词元，而是迭代地优化部分掩码序列 [13], [14] ，并可在解决更可靠位置的问题时，暂缓对不确定位置作出决定 [15]。图 1 显示，这种差异会直接影响硬件表现：在权重扰动下，尽管 Llama-3-8B 的预训练词元数更多，LLaDA-8B 在两个基准测试上的性能退化都平缓得多，并且在中等噪声水平下于 GSM8K 上超过 Llama-3。因此，这种鲁棒性是一种独特的生成特性，而非训练规模所致。

图2展示了这一行为的一种可能机制。在自回归解码中，每个词元一生成就会被确定下来：由噪声引发的翻转会使 *water* 立即变成*wine* ，并立刻进入因果上下文，引发连锁反应。相比之下，掩码扩散语言模型会先确定置信度高的位置，同时保留对不确定位置进行修改的可能，因此，最初偏向错误词元的扰动可以在更多上下文得到解析后重新评估，而恢复的预测间隔则会阻止翻转。然而，事先并不清楚观察到的鲁棒性究竟仅仅源于迭代去噪和轮间噪声平均，还是源于这些承诺与错误传播动态。回答这一问题对软硬件协同设计至关重要：如果延迟承诺是鲁棒性的来源，扩散语言模型就能容忍更大的持久性权重扰动，并降低对硬件保真度的要求。第III节中的分析将检验这些可能性，并且

支持基于承诺的解释。

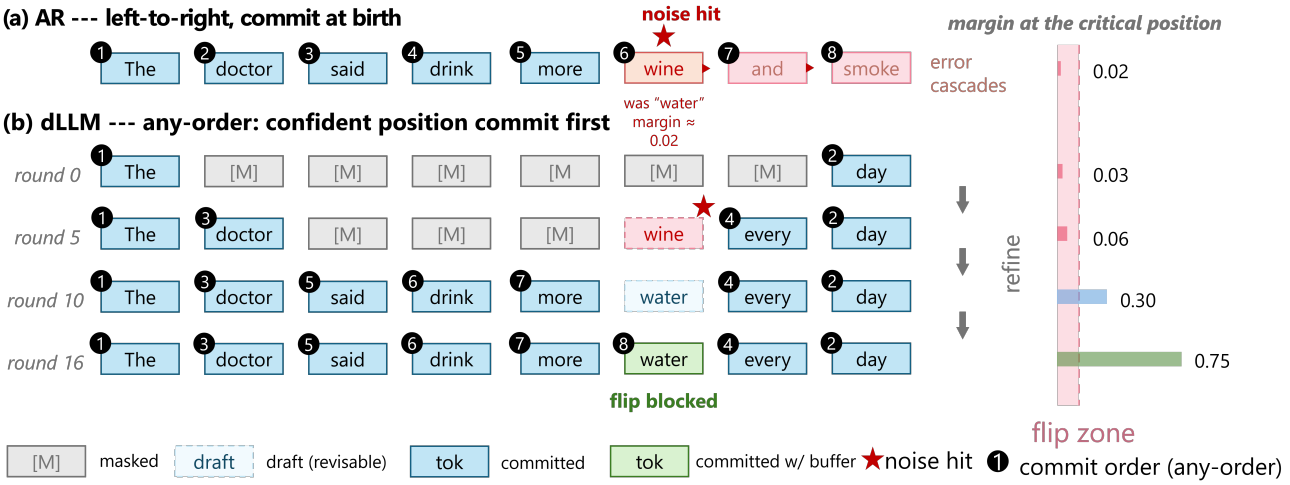


图2. 权重噪声下的承诺时序：这是图1中鲁棒行为的原因，由第III-B节和第III-C节中的受控分析所确定。自回归解码器（顶行）在每个词元生成时立即承诺，因此，早期由噪声引起的翻转（水→酒）会进入因果上下文，并在后续预测中级联传播。掩码扩散语言模型（底部各行）首先承诺高置信度位置，并将不确定位置保持为掩码或允许重新修订；随着更多上下文得到解析，关键位置的预测间隔得以恢复，从而阻止翻转。

在本研究中，我们研究了模拟存内计算引发的权重扰动下扩散语言模型鲁棒性，从受控分析逐步推进到校准仿真和板载测量。我们首先表明，在 PIQA [16]和 GSM8K [17]上，与规模匹配的模型和开放检查点相比，经过测试的扩散语言模型检查点相较于经过测试的自回归检查点能够更平稳地退化。随后，受控的承诺干预和上下文分析表明，延迟承诺，而非跨轮次噪声平均，是支持这一行为的一种机制。最后，我们通过芯片校准扰动模型以及在一个包含 16 个芯片的模拟存内计算集群上进行的混合测量，量化了由此产生的硬件优势。我们的贡献如下：

- 我们表明，在规模匹配的模型和下游基准测试中，扩散语言模型在AIMC引起的权重扰动下，相较于自回归模型能够更平缓地退化。
- 我们表明，这种鲁棒性并不能由跨轮次噪声平均来解释，并确定延迟词元承诺和渐进式上下文解析是得到支持的机制。
- 我们在一个16芯片28纳米电可擦除闪存模拟存内计算集群上展示了硬件优势：在根据片上矩阵向量乘法误差校准的扰动模型下，对八项基准测试进行8B规模扫描，并结合混合板载测量（变换器线性矩阵向量乘法在片上执行，其余部分采用数字方式），结果表明扩散语言模型可容忍大1.4–3×倍的权重噪声；增加去噪步数可以在实测延迟和能耗与更小的干净与噪声性能差距之间进行权衡，从而提供一种扩散语言模型特有的延迟—鲁棒性调节旋钮。

## II. 相关工作

**模拟存内计算**。模拟存内计算通过在存储器阵列中执行矩阵–向量乘法来加速神经网络推理，减少数据搬移并提高能效 [18], [19]。然而，器件变化、电导漂移、有限的编程精度，以及外围电路非理想性会扰动有效权重并降低推理准确率 [5], [6]。以往研究通过硬件感知训练、校准、补偿和抗噪架构来应对这些影响 [5], [7]–[20]。这些研究主要关注判别式工作负载或词元级准确率，而模拟存内计算扰动对现代语言生成中序列误差动态的影响仍未得到充分理解。

**自回归生成的鲁棒性**。自回归语言模型按顺序生成词元，每次预测都以上一步生成的词元为条件。这一解码过程可能通过暴露偏差和误差传播放大早期错误 [9], [10], [21]。现有方法通过改进训练目标、解码策略或生成后细化来减轻此类错误。相比之下，我们研究由不完善的推理硬件引入的错误：扰动引起的词元翻转可能被不可逆地插入因果上下文，导致后续预测以损坏序列为条件。

**扩散语言模型**。扩散语言模型通过迭代去噪生成文本，而非采用固定的从左到右解码。离散扩散方法和掩码扩散方法已展现出越来越有竞争力的语言建模和生成质量，同时支持并行或基于置信度的词元更新 [12]–[15], [22], [23]。近期研究还探索了改进扩散语言模型的采样、自我纠正和缩放行为 [24], [25]。这些工作着重于理想计算条件下的生成质量、采样效率或可控性。相比之下，我们研究了模拟存内计算引发的权重扰动对扩散语言模型的影响。我们表明，其优势并非源于跨轮次噪声平均，而是源于延迟词元承诺：不确定的词元在更可靠的上下文可用之前始终可以修订，从而抑制硬件诱导的错误级联。

III. 鲁棒性评估与分析

我们评估了扩散语言模型在持久性权重扰动下的鲁棒性，并将其行为与自回归模型进行比较。

## A. 权重扰动与鲁棒性评估

我们考虑推理时权重扰动。令 $\theta$ 表示干净模型参数，令 $\Delta\theta$ 表示每个扰动样本的数据噪声水平 $\sigma$。扰动后的参数为

$$\tilde{\theta} = \theta + \Delta\theta. \tag{1}$$

对于每个映射参数，经过校准的扰动遵循

$$\Delta\theta_{ij} \sim \mathcal{N}\left(0, (k_{\text{std}}|\theta_{ij}| + b_{\text{std}})^2\right). \tag{2}$$

GPU 实验采用权重主导极限 $b_{\text{std}} \to 0$，这等价于 $\tilde{\theta} = \theta \odot (1 + \sigma Z)$，其中 $Z_{ij} \sim \mathcal{N}$ 独立服从 (0, 1)。所有映射的线性层权重和偏置都会受到扰动；嵌入和语言模型头保持为数字形式，不在扰动范围内。该记法将加性物理模型与仿真中使用的有效乘性扰动统一起来。对于每项基准评估任务，每个噪声种子采样一张误差映射，并在该任务的每个样本和每个轮次中重复使用。因此，它表示一个编程硬件实例；采样器随机性与误差映射相互独立。扰动在所有自回归解码步骤或扩散语言模型去噪轮次中保持不变：

$$\tilde{\theta}^{(r)} = \tilde{\theta}, \qquad \forall r, \tag{3}$$

其中，$r$ 表示一个自回归解码步骤或扩散语言模型的一轮去噪。此设置模拟持久编程和器件失配，从而隔离固定编程误差图对完整生成轨迹的影响。

GSM8K 采用灵活的数值提取方式；LLaDA 和 Llama 分别使用 5-shot 和 8-shot 提示，并遵循其已发布的评估协议；下文的匹配模型比较则通过完全对齐的训练与评估设置，为这项开放检查点研究提供补充。主要噪声扫描覆盖完整的 GSM8K 测试集（1,319 道题），包含七个噪声水平和三个噪声种子；PIQA 研究使用完整的验证集划分（1,838 个样例）。工作区包含样本级输出、可用时的三至五个噪声种子、匹配的小型模型的四个训练种子，以及 scripts/ 和 training_causality/ 中的聚合脚本，为所有报告的趋势提供种子级分辨率。

对于性能指标 $A$，我们将模型 $M$ 在扰动水平 $\sigma$ 下的噪声诱导退化定义为

$$D_M(\sigma) = A_M(0) - A_M(\sigma), \tag{4}$$

其中，$A_M(0)$ 和 $A_M(\sigma)$ 分别表示干净和受扰动后的性能。较小的值 $D_M(\sigma)$ 表明对权重扰动具有更强的鲁棒性。在评估验证损失 $\mathcal{L}$ 时，我们则报告相对损失增幅

$$R_M(\sigma) = \frac{\mathcal{L}_M(\sigma) - \mathcal{L}_M(0)}{\mathcal{L}_M(0)}. \tag{5}$$

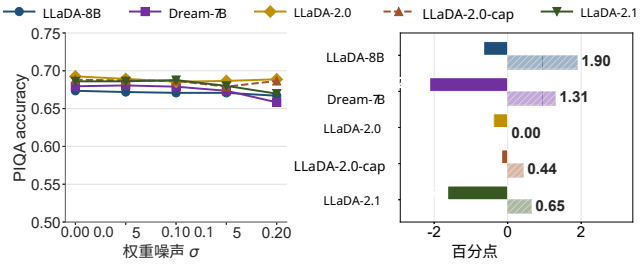


图 3. 开放式扩散语言模型检查点在权重扰动下的 PIQA 鲁棒性。**左：**准确率随扰动强度 $\sigma$ 的变化。**右：**在 $\sigma = 0.2$下的准确率差异，分解为干净权重下已存在的差异，以及扰动导致的额外性能下降。

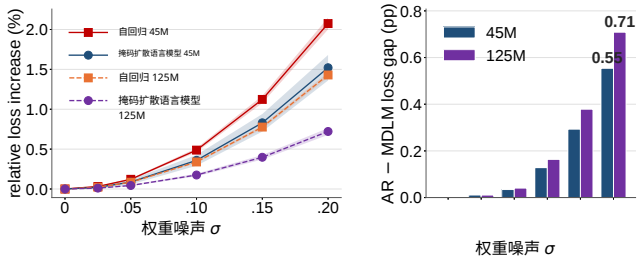


图 4. 匹配的自回归模型与掩码扩散语言模型在权重扰动下的受控比较。**左：**相对验证损失增幅随扰动强度 $\sigma$ 的变化。**右：**自回归模型相对于掩码扩散语言模型的额外相对损失增幅。

我们首先在 PIQA 上评估多个开源扩散语言模型检查点。图 3 展示了不同扰动强度下的准确率，并将噪声模型之间的差异分解为干净模型差距和扰动导致的性能下降，见 $\sigma = 0.2$。

图 3 中的分解很重要，因为带噪声模型之间的性能差距本身并不能证明鲁棒性存在差异：其中一部分差距可能在无噪声设置下就已存在。因此，我们区分受扰动后的绝对性能与噪声导致的性能下降。

接下来，我们对规模匹配的自回归模型和掩码扩散语言模型进行受控比较。由于这两个模型在规模上匹配，并在一致的训练和扰动设置下进行评估，因此该比较能够隔离出不能仅归因于模型规模或无噪声模型质量的鲁棒性差异。图 4 报告了公式（5）中定义的相对验证损失增幅。

掩码扩散语言模型相比其自回归对应模型表现出更小的噪声引起的相对损失增幅。结合开放检查点评估结果，这一结果表明，所观察到的鲁棒性并不仅仅是某个特定模型检查点的结果，也不是干净性能方面的优势。

## B. 受控生成协议

对扩散语言模型鲁棒性的一种可能替代解释是，反复进行的去噪轮次会平均掉独立的参数扰动。这一机制并不适用于公式（3）中的默认噪声设置，因为在整个轨迹中始终使用相同的受扰动

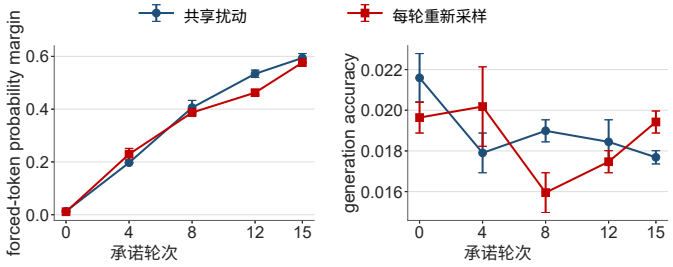


图 5. DLM 生成过程中扰动调度的影响。在共享噪声条件下，整个去噪轮次中保留同一个权重扰动；在重采样噪声条件下，每轮都会独立重新抽取扰动。**左侧：** 各承诺轮次的强制令牌边际。**右侧：** 各承诺轮次的生成准确率。

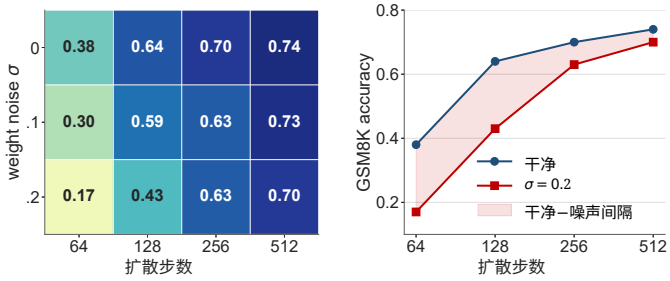


图 6. 不同扩散步数预算和权重扰动强度下的 GSM8K准确率 $K$ $\sigma$。增加去噪预算可缩小干净权重与噪声权重之间的性能差距。

模型。尽管如此，我们仍通过比较共享扰动与按轮次重新采样的扰动，对这一解释进行了明确检验。

共享噪声条件遵循公式 (3)；重采样噪声条件则在每轮抽取独立噪声扰动：

$$\tilde{\theta}^{(r)} = \theta + \Delta\theta^{(r)}, \qquad \Delta\theta^{(r)} \sim \mathcal{P}_\sigma, \tag{6}$$

其中，$\mathcal{P}_\sigma$ 表示噪声水平为 $\sigma$ 时的扰动分布。如果鲁棒性主要源于跨轮次对独立噪声进行平均，那么重采样噪声条件应当比共享噪声带来系统性优势。

图 5 比较了两种调度方案在各承诺轮次中的强制词元边际和生成准确率。

逐轮重新采样并未显著优于共享噪声条件。这排除了将独立的逐轮平均视为充分解释的可能性，但仅凭这一点无法确定唯一的作用机制。下文的承诺时间干预将直接检验改变承诺时间是否会改变干预后发散。

我们进一步改变扩散步数预算 $K$，以确定增加去噪轮次是否能专门提升鲁棒性。对于每种步数预算和噪声水平，我们测量干净与含噪性能差距

$$D(K, \sigma) = A(K, 0) - A(K, \sigma). \tag{7}$$

这一指标将干净生成性能的一般性提升与对含噪权重敏感性的降低区分开来。图 6 展示了不同扩散步数预算和扰动水平下的 GSM8K准确率。

随着扩散步数预算增加，干净性能与扰动性能之间的差距缩小。因此，增加去噪轮次不仅能提升干净生成质量，还能在最终确定词元决策之前，创造更多机会来修正或推迟不确定预测。

### C. 上下文解析分析

为揭示这一行为背后的机制，我们分析权重噪声下已解析上下文如何影响词元级置信度

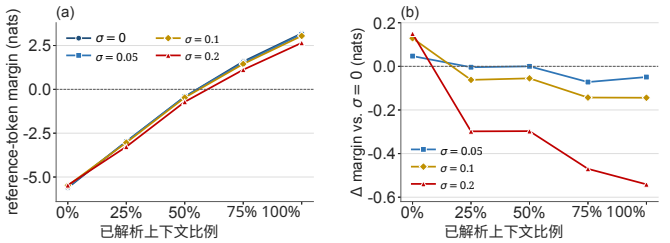


图 7。权重扰动下的上下文解析分析。**左：** 参考词元的对数概率边际随已解析上下文比例的变化。**右：** 相对于干净条件，权重扰动所引起的边际变化（$\sigma = 0$）。

。对于目标位置 $i$，令 $x_i^\star$ 表示参考词元，并令 $\mathcal{C}_i(q)$ 表示当相关位置中有 $q$ 的比例已解析时可用的上下文。我们将参考词元的对数概率裕度定义为

$$m_i(q, \sigma) = \log p_{\tilde{\theta}}\left(x_i^\star \mid \mathcal{C}_i(q)\right) - \max_{v \neq x_i^\star} \log p_{\tilde{\theta}}\left(v \mid \mathcal{C}_i(q)\right). \tag{8}$$

正边际表示参考词元的排名高于所有其他词元。相反，接近于零的边际表示决策容易发生翻转，此时微小扰动可能会改变排名最高的词元。

为隔离权重噪声的局部影响，我们相对于干净权重计算边际变化：

$$\Delta_{\text{noise}}(q, \sigma) = m(q, \sigma) - m(q, 0). \tag{9}$$

负值表示扰动会降低对参考词元的置信度。我们将这一惩罚与随着更多上下文得到解析而获得的边际增益进行比较：

$$G_{\text{ctx}}(\sigma) = m(q_{\max}, \sigma) - m(q_{\min}, \sigma). \tag{10}$$

图 7 比较了上下文解析的影响与权重扰动的直接局部影响。

随着解析出更多预言机上下文，无论权重干净与否，参考词元的间隔都会单调增加。相比之下，权重噪声导致的边际变化相对较小，观察到的最大变化为 $-0.541$ 纳特。因此，解析出的上下文带来的置信度提升，远大于扰动导致的局部置信度损失。

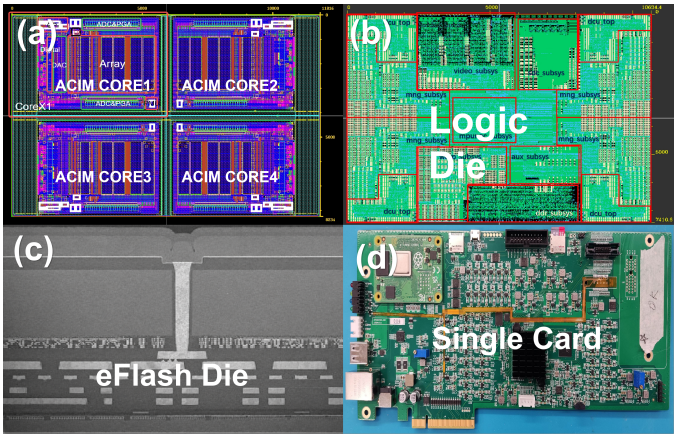


图 8. 基于28纳米嵌入式闪存的模拟存内计算原型。(a) 模拟存内计算裸片，包含四个核心，每个核心均集成了配备数模转换器和模数转换器/可编程增益放大器外设的权重驻留阵列。(b) 逻辑裸片布局，集成了 PCIe/DDR、视频、管理以及微处理器/片上网络子系统。(c) 混合键合三维堆叠的透射电子显微镜截面图（12纳米逻辑裸片、硅通孔层和28纳米嵌入式闪存裸片）。(d) 组装完成的单卡原型。

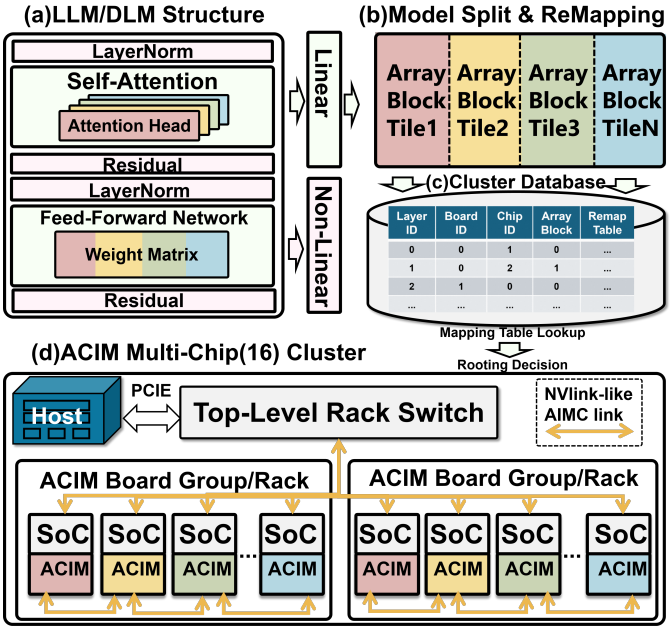


图 9. 模型在模拟存内计算集群上的部署。（a）作为映射目标的线性层。（b）围绕有缺陷列进行阵列块平铺和重新映射。（c）用于层到电路板路由的集群数据库。（d）配备类NVLink模拟存内计算链路和PCIe主机的16芯片集群。

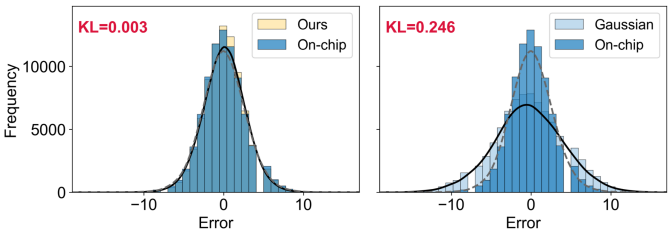


图 10. 经过实测片上矩阵向量乘法误差校准的权重扰动模型（KL ＝0.003），与未经校准的高斯分布（KL ＝0.246）的对比。

由于这项实验是从真实序列中揭示参考词元，因此它检验的是局部间隔假设，而非完整的自然展开。在一项互为补充的4500万参数研究中，自然展开上下文产生的强制词元间隔与预言机上下文相当或更大（比例为8时分别为0.30和0.30；比例为12时分别为0.56和0.33），而随机上下文则使间隔降至0.02。结合承诺干预实验，这些结果与延迟承诺和渐进式上下文解析能够减少误差传播的观点相符；但它们并未证明每次自然展开都具有预言机质量的上下文。

配对干预在固定提示、噪声种子和总去噪预算的同时，改变强制提交轮次。将承诺从第 0 轮推迟至第 8 轮后，4500万参数模型的强制后失配从 0.835 降至 0.265，1.25亿参数模型则从 0.85 降至 0.30。这表明承诺时序会影响误差传播，但并不能据此认定它是鲁棒性的唯一来源。

## 四、利用模拟存内计算上的扩散语言模型鲁棒性

### *A.* 硬件平台与集群部署

我们的硬件平台是一个基于28纳米嵌入式闪存的模拟存内计算原型（图 8），包含四个嵌入式闪存计算核心、数模/模数转换器外设，以及一个用于控制和PCIe/DDR的12纳米逻辑裸片，并通过混合键合集成为三维堆叠。我们在混合测量流程中使用16芯片集群；表 I定义了其计算边界。持久编程和器件失配促成了公式（1）中的静态扰动设定。

图 9 展示了如何将模型部署到由此类电路板组成的集群上。每个 Transformer块的线性算子（注意力投影和前馈网络权重矩阵）是映射目标：每个权重矩阵都被划分为阵列块瓦片，并通过重映射步骤，利用备用列将权重从故障阵列列周围重新定位。最终的分配——层标识符到（电路板、芯片、阵列块）的映射——连同重映射表一起记录在集群数据库中，运行时会查询该数据库，以便在整台机器中路由激活值。16 个芯片被组织成 SoC+模拟存内计算电路板，并置于顶层机架交换机之下。部署图描述的是已测量线性 MVM 的路由；不应将其解读为完整模型在芯片裸片上执行的证据。

### *B.* 校准噪声模型

校准遵循 NoiseGuard [20]，采用公式 (2) 中与权重相关的项以及偏置和 TIA/ADC 项；第 III-A 节定义了 GPU 简化和参数范围。拟合得到的边缘 MVM 误差直方图的 KL 散度为 0.003，而未经校准的高斯分布为 0.246（图 10），82× 这一降幅构成了整个混合流程所用校准扰动模型的依据。混合工作点报告为 $\sigma_{\text{eff}} = 0.15$。

图11使用该校准模型，在 8×L20 GPU 服务器上通过仿真，对大规模扩散语言模型和自回归模型进行比较。两种扩散语言模型（LLaDA-8B、Dream-7B）和两种自回归模型

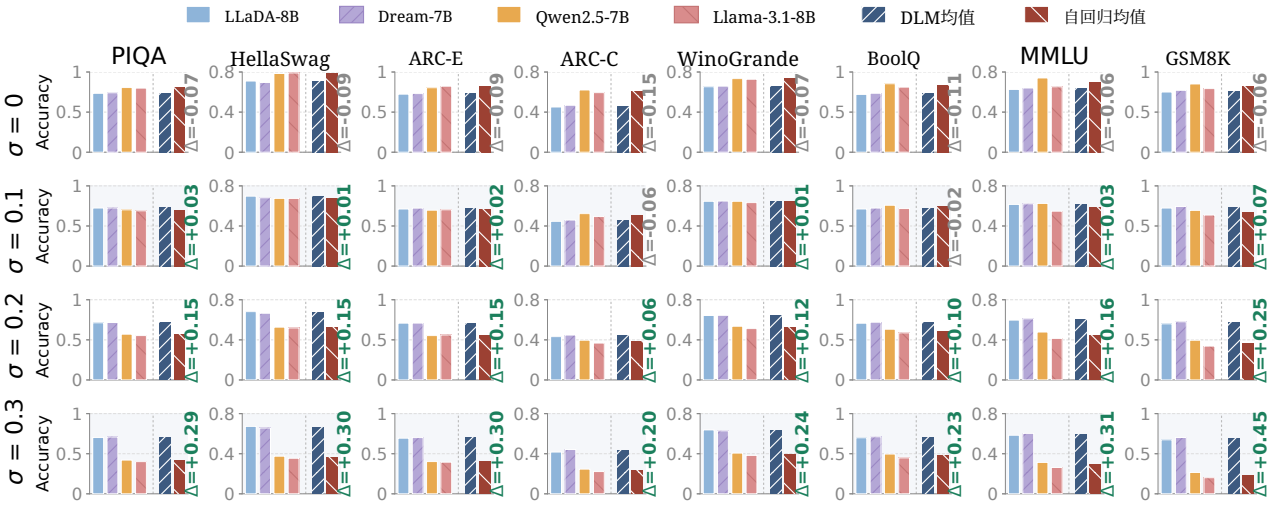


图11。在 8×L20 GPU服务器上进行校准权重噪声扫描。各行扫描 $\sigma \in \{0, 0.1, 0.2, 0.3\}$；柱状图报告两个扩散语言模型（LLaDA-8B、Dream-7B）、两个自回归模型（Qwen2.5-7B、Llama-3.1-8B）及组均值。 Δ：扩散语言模型均值优势，在全部八项基准测试上于 $\sigma = 0.3$ 达到 +0.20–+0.45 。

表 I $\sigma_{\text{eff}}$=0.15 下的混合质量（扩散语言模型：256步）。注意力/前馈神经网络矩阵-向量乘法：模拟存内计算；其他计算：数字。45M/125M：损失（↓）；8B：GSM8K准确率（↑）。最大 $\sigma_{\text{eff}}$：暂定2% 相对损失阈值

| 规模 | 模型 | 质量 | | $\Delta_{\text{rel}}$ | 最大值 $\sigma_{\text{eff}}$ |
|---|---|---|---|---|---|
| | | 数字 | 片上 | | |
| *45M* | AR | 3.412 | 3.451 | +1.14% | 0.19 |
| | MDLM | 3.465 | 3.494 | +0.84% | 0.27 |
| *125M* | AR | 3.183 | 3.207 | +0.75% | 0.28 |
| | MDLM | 3.236 | 3.242 | +0.19% | 0.58 |
| *8B*（开放） | Llama-3 | 0.748 | 0.512 | −31.6% | 0.03 |
| | LLaDA | 0.702 | 0.628 | −10.5% | 0.09 |

（Qwen2.5-7B、Llama-3.1-8B）在扰动强度为 $\sigma \in \{0, 0.1, 0.2, 0.3\}$的条件下，针对八项基准测试进行评估。自回归模型在 $\sigma = 0$时的初始干净准确率较高，但其优势在 $\sigma = 0.1$ 时就已消失，并在更大的扰动下反转：在 $\sigma = 0.3$ 时，DLM均值在每项基准测试上都超过自回归模型均值，差距从20个百分点（ARC-C）到45个百分点（GSM8K）不等。这一结果将观测到的趋势扩展到了大规模校准仿真中。

### *C.* 电路板评估

除非另有说明，我们使用图 9 中的映射流程，在 $\sigma_{\text{eff}} = 0.15$ 下评估 256 个 DLM 步骤。表 I 显示，在两个匹配规模下，MDLM 的性能下降幅度都更小；在 125M 规模下，损失增加了 0.19%，而 AR 增加了 0.75%。阈值条目给出了有利于 DLM 的 1.4–3× 配对比率，该比率从 45M 规模下的 1.4× 扩大到 8B 规模下的 3× 。

表 II 展示了扩散语言模型的质量与成本权衡：随着步数从 64 增至 512，LLaDA 的相对质量下降幅度从 40.4% 降至 3.9%

表 II $\sigma_{\text{eff}}$=0.15下的步骤–质量–成本权衡，指标和混合边界与表 I 相同。

| 模型 | 步骤 | 质量 | | | 成本/生成 | |
|---|---|---|---|---|---|---|
| | | 数字 | 片上 | $\Delta_{\text{rel}}$ | 延迟（秒） | 能耗（毫焦） |
| *LLaDA-8B* | 64 | 0.381 | 0.227 | −40.4% | 0.73 | 410 |
| | 128 | 0.642 | 0.503 | −21.7% | 1.42 | 801 |
| | 256 | 0.702 | 0.628 | −10.5% | 2.81 | 1590 |
| | 512 | 0.741 | 0.712 | −3.9% | 5.58 | 3150 |
| 掩码扩散语言模型-*125M* | 32 | 3.251 | 3.297 | +1.42% | 0.06 | 28 |
| | 64 | 3.243 | 3.268 | +0.77% | 0.11 | 54 |
| | 128 | 3.238 | 3.250 | +0.37% | 0.22 | 107 |
| | 256 | 3.236 | 3.242 | +0.19% | 0.43 | 211 |

报告的成本近似线性增长。标准自回归解码会立即确定每个词元，且无法控制去噪步数；因此，步数预算是扩散语言模型特有的部署调节参数，可在混合测量边界内权衡生成成本与抗噪能力。

## 五、结论

我们研究了模拟存内计算权重扰动下的扩散语言模型。由于采用延迟词元承诺和渐进式上下文解析，扩散语言模型比自回归模型的性能退化更加平缓，而这并非源于跨轮次噪声平均。去噪预算是扩散语言模型特有的延迟—鲁棒性调节旋钮。混合式 16 芯片模拟存内计算集群的测量结果表明，扩散语言模型能够承受高出 1.4–3× 倍的权重噪声，证明扩散生成适用于抗噪模拟存内计算推理。

## 致谢

本研究部分得到北京自然科学基金项目（编号 L2607008）资助，部分得到北京新星计划项目（编号 20250484807）资助。

参考文献


[1] T. Brown、B. Mann、N. Ryder、M. Subbiah、J. D. Kaplan、P. Dhariwal、A. Neelakantan、P. Shyam、G. Sastry、A. Askell、等，“语言模型是少样本学习器”，神经信息处理系统进展，第33卷，1877–1901页，2020年。[2] R. Pope、S. Douglas、A. Chowdhery、J. Devlin、J. Bradbury、J. Heek、K. Xiao、S. Agrawal 和 J. Dean，“高效扩展Transformer推理”，载于第六届机器学习与系统会议论文集，*MLSys 2023*，美国佛罗里达州迈阿密，*2023*年*6*月*4*日至*8*日 （D. Song、M. Carbin 和 T. Chen 编），mlsys.org，2023年。[3] J. Fernandez、C. Na、V. Tiwari、Y. Bisk、S. Luccioni 和 E. Strubell，“大语言模型推理的能耗考量与效率优化”，载于计算语言学协会第*63*届年会论文集（第*1*卷：长篇论文）*ACL 2025*，奥地利维也纳，*2025*年*7*月*27*日至*8*月*1*日 （W. Che、J. Nabende、E. Shutova 和 M. T. Pilehvar 编），32556–32569页，计算语言学协会，2025年。[4] N. Leroux、P. Manea、C. Sudarshan、J. Finkbeiner、S. Siegel、J. P. Strachan 和 E. Neftci，“用于快速且节能的大语言模型的模拟存内计算注意力机制”，《自然·计算科学》，第5卷，第9期，813–824页，2025年。[5] M. J. Rasch、C. Mackin、M. Le Gallo、A. Chen、A. Fasoli、F. Odermatt、N. Li、S. Nandakumar、P. Narayanan、H. Tsai、等，“利用基于存内计算的加速器，对大规模、多样化的深度学习推理工作负载进行硬件感知训练”，《自然·通讯》，第14卷，第1期，文章编号5282，2023年。[6] V. Joshi、M. Le Gallo、S. Haefeli、I. Boybat、S. R. Nandakumar、C. Piveteau、M. Dazzi、B. Rajendran、A. Sebastian 和 E. Eleftheriou，“利用计算相变存储器实现准确的深度神经网络推理”，《自然·通讯》*style>*，第*11*卷，第*1*期，文章编号*2473*，*2020*年。*A. Eldebiky*、*G. L. Zhang*、*G. Bocherer*、*B. Li* 和 *U. Schlichtmann*，*“Correctnet*：通过误差抑制与补偿增强神经网络模拟存内计算的鲁棒性”，载于欧洲设计、自动化与测试会议暨展览会，*DATE 2023*，比利时安特卫普，*2023*年*4*月*17*日至*19*日，1–6页，IEEE，2023年。[8] J. B¨uchel、I. Chalas、G. Acampa、A. Chen、O. Fagbohungbe、H. Tsai、K. E. Maghraoui、M. L. Gallo、A. Rahimi 和 A. Sebastian，“模拟基础模型”，载于神经信息处理系统进展第*38*卷：*2025*年神经信息处理系统年会，*NeurIPS 2025*，美国加利福尼亚州圣迭戈，*2025*年*12*月*2*日至7日／墨西哥墨西哥城，*2025*年*11*月*30*日至*12*月*5*日（D. Belgrave、C. Zhang、L. N. Montoya、H. Lin、R. Pascanu、P. Koniusz、M. Ghassemi、N. Chen、I. V. M. Ru´ız 和 A. Loaiza-Bonilla 编），2025年。[9] K. Arora、L. E. Asri、H. Bahuleyan 和 J. C. K. Cheung，“为何暴露偏差至关重要：从模仿学习视角看语言生成中的误差累积”，载于计算语言学协会论文集：*ACL 2022*，爱尔兰都柏林，*2022*年*5*月*22*日至*27*日（S. Muresan、P. Nakov 和 A. Villavicencio 编），*ACL*论文集第 ACL 2022 卷，700–710页，计算语言学协会，2022年。[10] M. Ranzato、S. Chopra、M. Auli 和 W. Zaremba，“循环神经网络的序列级训练”，载于第*4*届国际学习表征会议，*ICLR 2016*，波多黎各圣胡安，*2016*年*5*月*2*日至*4*日，会议论文集（Y. Bengio 和 Y. LeCun 编），2016年。[11] M. Zhang、O. Press、W. Merrill、A. Liu 和 N. A. Smith，“语言模型幻觉如何滚雪球式扩大”，载于第*41*届国际机器学习会议，*ICML 2024*，奥地利维也纳，*2024*年*7*月*21*日至*27*日（R. Salakhutdinov、Z. Kolter、K. A. Heller、A. Weller、N. Oliver、J. Scarlett 和 F. Berkenkamp 编），机器学习研究论文集第235卷，59670–59684页，PMLR／OpenReview.net，2024年。[12] S. Nie、F. Zhu、Z. You、X. Zhang、J. Ou、J. Hu、J. Zhou、Y. Lin、J. Wen 和 C. Li，“大语言扩散模型”，载于神经信息处理系统进展第*38*卷：*2025*年神经信息处理系统年会，*NeurIPS 2025*，美国加利福尼亚州圣迭戈，*2025*年*12*月*2*日至*7*日／墨西哥墨西哥城，*2025*年*11*月*30*日至*12*月*5*日（D. Belgrave、C. Zhang、L. N. Montoya、H. Lin、R. Pascanu、P. Koniusz、M. Ghassemi、N. Chen、I. V. M. Ru´ız 和 A. Loaiza-Bonilla 编），2025年。[13] J. Austin、D. D. Johnson、J. Ho、D. Tarlow 和 R. van den Berg，“离散状态空间中的结构化去噪扩散模型”，载于神经信息处理系统进展第34卷：2021年神经信息处理系统年会，NeurIPS 2021，12月 *6-14, 2021*，线上(M. Ranzato、A. Beygelzimer、Y. N. Dauphin、P. Liang 和 J. W. Vaughan 编)，第17981–17993页，2021年。[14] S. S. Sahoo、M. Arriola、Y. Schiff、A. Gokaslan、E. Marroquin、J. T. C hiu、A. Rush 和 V. Kuleshov，“简单有效的掩码扩散语言模型”，载于神经信息处理系统进展第*37*卷：*2024*年神经信息处理系统年会，*NeurIPS 2024*，温哥华，不列颠哥伦比亚省，加拿大，*2024*年*12*月*10*日至*15*日(A. Globersons、L. Mackey、D. Belgrave、A. Fan、U. Paquet、J. M. Tomczak 和 C. Zhang 编)，2024年。[15] J. Kim、K. Shah、V. Kontonis、S. M. Kakade 和 S. Chen，“为最坏情况训练，为最好情况规划：理解掩码扩散中的词元排序”，载于第*42*届国际机器学习会议，*ICML 2025*，温哥华，不列颠哥伦比亚省，加拿大，*2025*年*7*月*13*日至*19*日(A. Singh、M. Fazel、D. Hsu、S. Lacoste-Julien、F. Berkenkamp、T. Maharaj、K. Wagstaff 和 J. Zhu 编)，机器学习研究论文集第267卷，PMLR / OpenReview.net，2025年。[16] Y. Bisk、R. Zellers、R. L. Bras、J. Gao 和 Y. Choi，“PIQA：自然语言中的物理常识推理”，载于第三十四届*AAAI*人工智能会议，*AAAI 2020*；第三十二届人工智能创新应用会议，*IAAI 2020*；第十届*AAAI*人工智能教育进展研讨会，*EAAI 2020*，*2020*年*12*月，*7*7432–7439，2020美国纽约州纽约市，*2*月-第–页，AAAI出版社，。[17] K. Cobbe、V. Kosaraju、M. Bavarian、M. Chen、H. Jun、L. Kaiser、M. Plappert、J. Tworek、J. Hilton、R. Nakano、C. Hesse 和 J. Schulman，“训练验证器以解答数学应用题”，*CoRR*，第abs/2110.14168卷，2021年。[18] A. Shafiee、A. Nag、N. Muralimanohar、R. Balasubramonian、J. P. Strachan、M. Hu、R. S. Williams 和 V. Srikumar，“Isaac：一种利用交叉阵列中的原位模拟运算的卷积神经网络加速器”，载于*ACM SIGARCH*计算机体系结构新闻，第44卷，第14–26页，ACM，2016年。[19] Khaddam-Aljameh 等，“Hermes-core——采用14纳米互补金属氧化物半导体工艺、基于相变存储器的1.59 TOPS/mm²存内计算核心，使用300 ps/LSB线性化的基于CCO的模数转换器”，*IEEE*固态电路期刊，第57卷，第4期，第1027–1038页，2022年。[20] G. Wang 等，“NoiseGuard：面向存内计算片上系统的综合框架，涵盖噪声建模、噪声感知训练和噪声补偿”，*IEEE*计算机辅助设计与集成电路系统汇刊，第45卷，第6期，第2597–2610页，2026年。[21] S. Bengio、O. Vinyals、N. Jaitly 和 N. Shazeer，“循环神经网络序列预测的计划采样”，载于神经信息处理系统进展第*28*卷：*2015*年神经信息处理系统年会，*2015*年*12*月7日至*12*日，加拿大魁北克省蒙特利尔(C. Cortes、N. D. Lawrence、D. D. Lee、M. Sugiyama 和 R. Garnett 编)，第1171–1179页，2015年。[22] A. Lou、C. Meng 和 S. Ermon，“通过估计数据分布的比率进行离散扩散建模”，载于第*41*届国际机器学习会议，*ICML 2024*，奥地利维也纳，*2024*年*7*月*21*日至*27*日(R. Salakhutdinov、Z. Kolter、K. A. Heller、A. Weller、N. Oliver、J. Scarlett 和 F. Berkenkamp 编)，机器学习研究论文集第235卷，第32819–32848页，PMLR / OpenReview.net，2024年。[23]M. Ghazvininejad、O. Levy、Y. Liu 和 L. Zettlemoyer，“Mask-predict：条件掩码语言模型的并行解码”，载于*2019*年自然语言处理实证方法会议暨第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第6112–6121页，2019年。[24] G. Wang、Y. Schiff、S. S. Sahoo 和 V. Kuleshov，“通过推理时扩展对离散扩散模型重新掩码”，载于神经信息处理系统进展第*38*卷：*2025*年神经信息处理系统年会，*NeurIPS 2025*，美国加利福尼亚州圣迭戈，*2025*年*12*月*2*日至*7*日／墨西哥墨西哥城，*2025*年*11*月*30*日至*12*月*5*日(D. Belgrave、C. Zhang、L. N. Montoya、H. Lin、R. Pascanu、P. Koniusz、M. Ghassemi、N. Chen、I. V. M. Ru´ız 和 A. Loaiza-Bonilla 编)，2025年。[25] S. Nie、F. Zhu、C. Du、T. Pang、Q. Liu、G. Zeng、M. Lin 和 C. Li，“扩展文本上的掩码扩散模型”，载于第十三届国际学习表征会议，*ICLR 2025*，新加坡，*2025*年*4*月*24*日至*28*日，OpenReview.net，2025年。