用于 Transformer 权重压缩的正交混合符号编码与渐进式残差细化

Anindya Roy Chowdhury Atanu Barai 洛斯阿拉莫斯国家实验室 洛斯阿拉莫斯国家实验室achowdhury@lanl.gov abarai@lanl.gov

Nandakishore Santhi 洛斯阿拉莫斯国家实验室nsanthi@lanl.gov

arXiv:2611.31089v1 [cs.AI] 23 Nov 2026

摘要

摘要

大语言模型量化可减少存储需求，但在低比特率下往往会降低预测保真度。我们提出一种免训练、无 CDF、受高斯中心极限定理启发的渐进式残差符号量化框架，该框架采用随机正交混合，促使残差分布趋近高斯分布，并改善 Transformer 权重的基于符号的重建。每个阶段都会独立归一化残差，并通过解析校准重建结果，逐步恢复量化误差。该框架无需学习型码本、经验累积分布函数或网格优化，即可灵活权衡存储与保真度。我们在 GPT-2 Large 和 Llama-3-8B 的注意力层及多层感知机投影上进行了评估。在五个残差阶段下，联合量化能够保持接近基线的预测保真度，困惑度分别达到 32.16 和 12.73，而全精度值分别为 29.89 和 11.07。消融实验进一步表明，与基于恒等变换的符号编码相比，正交混合显著改善了重建效果；激活感知缩放能够提升输出保真度，而低精度统计量则可在几乎不增加困惑度的情况下减少存储开销。与传统量化器的比较证实了该方法能够逐步恢复保真度，并揭示了其存储–质量权衡和运行时权衡。结果表明，正交混合残差符号编码适用于灵活、免训练的 Transformer 权重压缩。

关键词： 大语言模型、训练后量化、正交混合、渐进式残差编码。

1 引言

Transformer 模型（Vaswani 等人，2017），包括 GPT-2 和 LLaMA 系列（Radford 等人，2019；Touvron 等人，2023a,b），会带来高昂的权重存储、传输和推理内存成本。训练后量化（PTQ）无需重新训练即可降低这些成本。成熟的方法包括感知离群值的 LLM.int8()（Dettmers 等人，2022）、二阶 GPTQ（Frantar 等人，2023）、感知激活的 SmoothQuant 和 AWQ（Xiao 等人，2023；Lin 等人，2024），以及基于旋转的

QuaRot 和 SpinQuant Ashkboos 等人（2024）；Liu 等人（2025）。然而，标量量化并不能直接利用权重向量各坐标之间的关系。

向量量化（VQ）通过联合编码来捕捉这种结构，在信息论和信号处理领域有着悠久的历史：给定一个由 n 个 d 维向量构成的矩阵 $W \in \mathbb{R}^{d \times n}$，将其映射到一个由代表性向量组成的离散有限集合，称为码本。随后，仅使用该码本即可近似重建原始矩阵。近年来，向量量化已成为人工智能基础设施的核心技术，相关方法包括采用加性码本的技术 Egiazarian 等人（2024）、数据感知码本和误差补偿 van Baalen 等人（2024）、结构化格点码本 Tseng 等人（2024a），以及网格编码 Tseng 等人（2024b）。这些方法引出了一个问题：能否在不使用学习型码本、网格搜索或依赖分布的累积分布函数（CDF）计算的情况下，对向量权重进行渐进式压缩，并明确控制重建保真度？这个问题很重要，因为固定的低比特表示可能无法提供足够的保真度，而存储额外的校正阶段则可能以增加存储成本为代价恢复准确率。

本文提出一种无需码本、无 CDF、在信息论意义上近乎最优的渐进式残差符号向量编解码器。在每个阶段，该编解码器都会对权重向量或残差向量进行中心化和均方根归一化，应用带种子的正交混合，并根据变换后坐标的符号进行编码。紧凑的统计量和拟合得到的最小二乘增益可用于重建；分块失真则用于在确定性变换候选项中进行选择。随后，编解码器将剩余误差编码为连续的符号编码校正，仅当某个阶段能够降低分块均方根失真时才予以接受。任何已接受的前缀都可解码，因此残差阶段深度成为明确的存储–保真度控制手段，无需针对每个工作点分别训练量化器。高斯中心极限定理行为为混合提供了动机，但正确性并不以此为前提，也不通过 CDF 或逆累积分布函数运算进行评估。激活校准不会更新模型权重；无需梯度、蒸馏或恢复微调。

我们在 GPT-2 Large 和 Llama-3-8B 上评估仅注意力、仅多层感知机以及注意力与多层感知机联合量化，测量不同残差深度和编解码器消融实验下的重建误差、困惑度、输出保真度、序列化存储和运行时间。我们的贡献包括：(i) 一种具有紧凑重建统计量、无 CDF 的基于符号的向量编解码器；(ii) 失真感知变换选择和单调残差细化；以及 (iii) 无需重新训练，对渐进式存储–保真度权衡进行模型级表征。

我们的结果表明，更深的残差细化能显著提升模型保真度，但也会增加存储需求和重建成本。在五个残差阶段下进行注意力与多层感知机联合量化时，GPT-2 Large 的困惑度为 32.16，Llama-3-8B 的困惑度为 12.73；相比之下，全精度基线的对应数值分别为 29.89 和 11.07。消融实验表明，激活感知缩放和变换选择有助于提升保真度。这些发现展现了免训练渐进式残差符号量化的潜力：通过调节残差阶段深度，在压缩效率与模型保真度之间取得平衡。

2 背景与相关工作

激进的权重压缩需要在有限的存储预算内保留有用的模型信息，包括重建所需的辅助信息。因此，表示方式至关重要：它决定了保留哪些信息、如何重建权重，以及增加额外比特如何改善近似效果。我们的工作研究了一种基于正交混合符号向量、可逐步细化的表示方式。其独特贡献在于对每个残差阶段进行设计，将基于失真的选择与重建统计量及变换元数据相结合。

研究人员提出了不同的训练后量化技术，用于在无需重新训练的情况下压缩 Transformer 权重矩阵$W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$。GPTQ 使用近似的二阶重建信息

Frantar 等人（2023）；AWQ 保护激活显著通道 Lin 等人（2024）；SmoothQuant 在权重和激活值之间转移量化难度 Xiao 等人（2023）；LLM.int8() 则单独处理离群值 Dettmers 等人（2022）。尽管这些方法贡献显著，但它们并未探索向量量化。

标量量化器对系数进行编码，而向量量化器对块进行编码。AQLM 使用加性码本（Egiazarian 等人（2024）），GPTVQ 使用数据感知向量编码（van Baalen 等人（2024））。QuIP# 将非相干性处理与结构化格点码本相结合（Tseng 等人（2024a）），而 QTIP 使用网格编码量化，在不采用指数规模码本的情况下实现较高的有效向量维度（Tseng 等人（2024b））。我们的编解码器则存储变换后的符号、重建统计量和变换元数据，形成一种既无需质心查找、也无需网格优化的块表示。

正交变换在保持欧几里得范数的同时重新分配坐标幅值。QuaRot 和 SpinQuant 使用旋转来提升量化鲁棒性（Ashkboos 等人（2024）；Liu 等人（2025））。TurboQuant 将随机旋转与分布匹配标量量化相结合，并在进行内积估计时采用一比特残差校正（Zandieh 等人（2026））。我们直接对正交混合后的向量进行符号编码，并递归地编码其残差。我们的编码器无需 *CDF*：既不需要经验累积分布函数估计，也不需要逆累积分布函数表。

残差量化通过 $\widehat{x}^{(r)} = \sum_{k=1}^{r} \widehat{e}^{(k)}$ 构建，并结合 $e^{(k)} = e^{(k-1)} - \widehat{e}^{(k)}$ 和 $e^{(0)} = x$。RRQ 对基底及其残差应用低比特标量量化器，从一个检查点生成多精度大语言模型权重 Luo 等人（2026）。我们的贡献是残差阶段编解码器：每个校正量都是按失真选择、经过正交混合的符号向量，根据（μ，σ，γ）和变换元数据进行重建。因此，区别在于阶段表示，而非残差细化本身：额外阶段通过符号向量几何细化近似，而不是提高标量精度。

LoRA 和 QLoRA 通过优化来调整模型权重 Hu 等人（2022）；Dettmers 等人（2023）。我们的量化器既不需要模型再训练，也不需要恢复适配器训练；校准样例仅用于提供激活统计量。这避免了额外的训练阶段，同时在编码过程中保留基于失真的选择。

随机正交混合会形成归一化块元素的加权和。中心极限定理论证表明，在满足适当正则性条件时，边缘分布近似为高斯分布，但这并不保证任意权重都具有高斯性。高斯性为混合提供了动机；编解码器本身既不需要进行高斯概率计算，也不需要使用经验累积分布函数映射。

3 问题表述

假设 $\mathcal{W}_{\mathcal{S}} = \{W_1, \ldots\ldots, W_L\}$ 表示在量化范围 $\mathcal{S}$ ∈{注意力、多层感知机、注意力 ＋多层感知机}下选定的权重，并令 $Q_r(\cdot)$ 表示具有 r 个阶段的渐进式残差符号量化。

$$\boxed{\widehat{W}_{\mathrm{Q},\ell}^{(r)} = Q_r(W_\ell),}$$

而未定向参数保持预训练状态。参数失真为

$$D_{\mathrm{param}}^{(r,\mathcal{S})} = \frac{\sum_{\ell \in \mathcal{S}} \|W_\ell - \widehat{W}_{\mathrm{Q},\ell}^{(r)}\|_F^2}{\sum_{\ell \in \mathcal{S}} |W_\ell|},$$

以均方根误差、相对误差和余弦相似度作为互补指标 .

对于输入 X，设 $Z_{\mathrm{FP}} = f_{\mathrm{FP}}(X)$ 和 $Z_{r,\,\mathcal{S}} = f_{r,\,\mathcal{S}}(X)$。函数失真在有效且未移位的词元位置上进行测量：

$$D_{\mathrm{logit}}^{(r,\mathcal{S})} = \frac{\sum_{i,t,v} a_{i,t}\big(Z_{\mathrm{FP},i,t,v} - Z_{r,\mathcal{S},i,t,v}\big)^2}{V \sum_{i,t} a_{i,t}}, \tag{1}$$

其中，$a_{i,t} \in \{0,1\}$ 用于掩码标记有效的 logit 位置，V为词汇表大小。与下文的下一词元目标掩码不同，$a_{i,t}$ 作用于未移位的词元位置。本文还报告 Logit 余弦相似度。语言模型质量通过掩码下一词元困惑度进行衡量，

$$\boxed{\mathrm{PPL} = \exp\left(-\frac{\sum_{i=1}^{N}\sum_{t=1}^{T_i-1} m_{i,t} \log p(x_{i,t+1} \mid x_{i,\le t})}{\sum_{i=1}^{N}\sum_{t=1}^{T_i-1} m_{i,t}}\right)}, \tag{2}$$

其中，$m_{i,t} \in \{0,1\}$ 用于标识评估样本 i 中有效的移位目标位置。参考模型和量化模型使用相同的留出样本、分词方式和掩码。

令 $S(r,\mathcal{S})$ 表示压缩存储，$D(r,\mathcal{S})$ 表示质量损失度量。我们通过评估已配置的残差深度和投影范围，刻画以下概念性的存储–质量工作点目标，而非求解约束优化问题：

$$\min_{r,\mathcal{S}} D(r,\mathcal{S}) \qquad \text{s.t.} \quad S(r,\mathcal{S}) \le B, \tag{3}$$

$$\min_{r,\mathcal{S}} S(r,\mathcal{S}) \qquad \text{s.t.} \quad \mathrm{PPL}(r,\mathcal{S}) \le P_{\max}. \tag{4}$$

因此，r 控制 编解码器容量以及$\mathcal{S}$ 压缩的 Transformer 权重类型和比例 .

4 所提方法

4.1 概述与激活感知校准

免训练流水线选择注意力、多层感知机或注意力+多层感知机投影，收集校准激活值，应用可选的激活感知等效缩放，并使用无 CDF 残差符号编解码器对选定权重进行分块编码。不对模型参数进行优化。

对于输入通道激活统计量 a_j 和参考幅值 a_{ref}，我们定义

$$s_j = \left(\frac{a_j}{a_{\mathrm{ref}}}\right)^{\beta}, \qquad S = \mathrm{diag}(s_1, \ldots, s_{d_{\mathrm{in}}}), \tag{5}$$

在满足所配置边界的条件下。等效变换写作

$$W' = WS, \tag{6}$$

$$x' = S^{-1}x, \tag{7}$$

$$W'x' = Wx, \tag{8}$$

这会改变呈现给编解码器的权重分布，而不改变全精度运算。

4.2 感知 CDF 的量化

当已知累积分布函数（CDF） F_X 时， b 位量化器可以将 X 映射为均匀变量 $U = F_X（X）$，并分配

$$q(X) = \min\left\{2^b - 1, \left\lfloor 2^b F_X(X) \right\rfloor\right\}. \tag{9}$$

重建使用逆累积分布函数：

$$\widehat{X} = F_X^{-1}\left(\frac{q(X) + 1/2}{2^b}\right). \tag{10}$$

这会生成等概率分箱，但不一定能最小化均方误差。此外，它还需要了解或估计 CDF。相比之下，我们的框架通过正交混合、符号编码和渐进式残差重建，避免了显式的 CDF 估计。

4.3 无累积分布函数残差符号编解码器

现在，我们消除了了解 CDF 的需要。对于行块中的每个长度为 n 的残差向量 e，编解码器会计算其均值和中心化均方根：

$$\mu = \frac{1}{n}\sum_{i=1}^{n} e_i, \qquad \sigma = \sqrt{\frac{1}{n}\|e - \mu\mathbf{1}\|_2^2}. \tag{11}$$

使用 $q \in \{16，32\}$ 位的存储统计量，令 $\widetilde{\mu} = Q_q(\mu)$ 和 $\widetilde{\sigma} = Q_q(\max\{\sigma，\epsilon\})$，其中 Q_q 表示舍入到指定格式。残差经过归一化、正交混合和符号编码：

$$z = \frac{e - \widetilde{\mu}\mathbf{1}}{\max\{\widetilde{\sigma}, \epsilon\}}, \tag{12}$$

$$u = Vz, \qquad VV^T = I_m, \tag{13}$$

$$s_i = \begin{cases} +1, & u_i \geq 0, \\ -1, & u_i < 0. \end{cases} \tag{14}$$

仅存储前 $m_{\text{eff}} \leq m$ 个符号；重建时，其余变换后坐标被置零。默认情况下， $m_{\text{eff}} = m = n$。

高斯中心极限定理动机。 在适当的弱依赖和非支配条件下，每个混合坐标 $u_i = \sum_j V_{ij} z_j$ 可能呈现近似高斯分布的行为。这为正交混合提供了动机，但既不保证高斯性，也不施加分布假设。该变换由带随机种子的高斯随机矩阵经过 QR 正交归一化生成。不同于采用网格编码量化的 QTIP Tseng 等人（2024b），我们的编解码器直接对变换后的符号进行编码，无需进行高斯累积分布函数或逆累积分布函数计算。

对于重建，令 $\bar{s} \in \mathbb{R}^m$ 包含保留的符号，后跟零值。定义

$$b = \widetilde{\sigma} V^T \bar{s}, \qquad t = e - \widetilde{\mu}\mathbf{1}. \tag{15}$$

以精度 q 拟合并存储每个向量、每个阶段的最小二乘增益：

$$\gamma_{\text{LS}} = \frac{t^T b}{\max\{b^T b, \epsilon\}}, \tag{16}$$

$$\widetilde{\gamma} = Q_q(\gamma_{\text{LS}}), \tag{17}$$

$$\hat{e} = \widetilde{\mu}\mathbf{1} + \widetilde{\gamma} b. \tag{18}$$

非有限有限增益会在舍入前替换为高斯派生的回退值 $\sqrt{2/\pi}$.

对于每个行块 B，编解码器会评估 K 个确定性候选变换并选择 s

$$j^* = \arg\min_{1\le j\le K} \frac{\|B - \widehat{B}(V_j)\|_F^2}{|B|}, \tag{19}$$

其中， $\widehat{B}(V_j)$包括渐进式残差细化和任何提前终止。仅保留选定的变换标识符、编码符号和统计量（$\tilde{\mu}$，$\tilde{\sigma}$，$\tilde{\gamma}$）。不需要经验累积分布函数表、学习型码本或网格路径。

4.4 逐次残差细化

逐次残差阶段量化通过符号编码校正逐步编码重建误差，从而提高权重保真度。其渐进式表示以及在每个层级上的近似最优性，使得推理时能够选择解码深度，在无需重新训练或重新量化的情况下，提供灵活的存储–保真度权衡。

对于每个行块 B，初始化 $E^{(0)} = B$ 和 $\widehat{B}^{(0)} = 0$。在阶段 k，对 $E^{(k-1)}$ 进行逐向量编码和解码会得到校正量 $\widehat{E}^{(k)}$ 并

$$\widehat{B}^{(k)}_{\text{cand}} = \widehat{B}^{(k-1)} + \widehat{E}^{(k)}, \tag{20}$$

$$E^{(k)}_{\text{cand}} = B - \widehat{B}^{(k)}_{\text{cand}}. \tag{21}$$

当 $\text{RMS}(A) = \sqrt{\|A\|_F^2/|A|}$时，单调细化仅在满足以下条件时接受一个阶段：

$$\text{RMS}(E^{(k-1)}) - \text{RMS}(E^{(k)}_{\text{cand}}) > \delta, \tag{22}$$

其中 $\delta \ge 0$ （默认值为 0）。接受的候选项将成为 $\widehat{B}^{(k)}$ 和 $E^{(k)}$；否则，块编码将终止。对于 $r_{\text{enc}} \le r$ 个接受的阶段，

$$\widehat{B}^{(r_{\text{enc}})} = \sum_{k=1}^{r_{\text{enc}}} \widehat{E}^{(k)}, \qquad E^{(r_{\text{enc}})} = B - \widehat{B}^{(r_{\text{enc}})}. \tag{23}$$

任何已接受的前缀均可解码，且接受与否按块而非按向量评估。

所提渐进式中心极限定理矢量量化的近似最优性。 对于独立同分布输入权重，经过正交混合的归一化残差坐标是独立同分布的 $u_i \sim \mathcal{N}(0, 1)$。对于恒振幅符号重建 $\hat{u}_i = c\,\text{sign}(u_i)$，其期望均方误差为

$$D(c) = 1 - 2c\sqrt{\frac{2}{\pi}} + c^2. \tag{24}$$

最优振幅为 $c^* = \sqrt{2/\pi}$，得到 $D(c^*) = 1 - 2/\pi \approx 0.3634$，在此重建族中，期望平方误差降低了 63.7%。

假设输入独立同分布且采用中心极限定理混合，所有连续阶段的残差也都恰好服从独立同分布的高斯分布，因此，渐进细化会按如下方式累积失真：

$$\frac{D_k}{n} = \left(1 - \frac{2}{\pi n}\right) D_{k-1}, \; \frac{D_L}{n} = \left(1 - \frac{2}{\pi n}\right)^L D_0. \tag{25}$$

作为 $n \gg 1$，$(1 - 1/n)^n \to e^{-1}$ 给出 $\frac{D_L}{n} \approx D_0 2^{-0.9185\frac{L}{n}}$

另一方面，对于无记忆（独立同分布）的高斯源，香农的率失真函数最差，给出的信息论失真下界为：

$$\frac{D_{\mathrm{RD}}(R)}{n} \geq D_0 2^{-2R}, \tag{26}$$

其中码率 $R = \frac{L}{n}$ 表示每坐标比特数。因此，预测的衰减指数 0.9185 约为最优高斯率失真指数的 45%，同时仍能随 L 实现指数级改进。

我们的实现采用最小二乘重建增益和分块单调接受机制，以确保实际块失真不会增加。然而，如果输入权重并非独立同分布，残差在各阶段就无法自动保证呈高斯分布——不过，由于中心极限定理在相对宽松的条件下适用，可以通过对输入权重进行归一化和白化来解决这一问题，但已存储的统计元数据可能会提高有效码率。

4.5 分块执行与存储

选定的矩阵被划分为彼此独立的行块，并附带相应的范围、投影类型、残差流、统计量和变换元数据。对于 P 编解码设备，块分配如下：

$$B_i \mapsto G_{\,1+((i-1) \bmod P)}, \tag{27}$$

从而支持多 GPU 量化。压缩块可按需重建，并可选择启用缓存；同时，这两种模型均可在单个加速器上评估。

检查点可能会将压缩投影和元数据作为叠加层存储，并从预训练模型中加载未定向参数。因此，我们分别报告选定的权重载荷、序列化叠加大小和完整的部署存储，因为叠加层大小并不代表一个自包含的压缩模型。

算法 1 无CDF渐进式残差符号量化

要求： 模型 f、范围$\mathcal{S}$、校准数据、阶段 r、候选项 K、统计位数 q、阈值 δ

选择投影并收集激活统计量
计算缩放因子并形成 $W' = WS$
对 每个选定的矩阵 W' **执行**
  将 W' 划分为行块 B
  对于 每个块 B **执行**
    生成变换$\{V_j\}_{j=1}^{K}$
    对于 $j = 1, \ldots, K$ **执行**
      $E \leftarrow B$, $\widehat{B}_j \leftarrow 0$
      对于 $k = 1, \ldots, r$ **执行**
        $\widehat{E} \leftarrow$ 符号编码解码(E, V_j, q)
        $E_{\text{new}} \leftarrow E - \widehat{E}$
        如果单调性且均方根(E) –均方根$(E_{\text{new}}) \le \delta$ **则**
          跳出
        结束条件判断
        $\widehat{B}_j \leftarrow \widehat{B}_j + \widehat{E}$
        $E \leftarrow E_{\text{new}}$
      结束循环
      $D_j \leftarrow \|B - \widehat{B}_j\|_F^2 / |B|$
    结束循环
    $j^* \leftarrow \arg\min_j D_j$
    使用 V_{j^*}重新编码 B
    打包接受的阶段和元数据
  结束循环
结束循环
保存压缩后的叠加层和流式处理模块

5 实验方法

我们在注意力、多层感知机以及注意力+多层感知机量化设置下评估 GPT-2 Large (`gpt2-large`) 和 Llama-3-8B (`meta-llama/Meta-Llama-3-8B`)，不包括嵌入层和语言模型头。每个范围使用 $r \in \{1, 2, 3, 5\}$，产生 $2 \times 3 \times 4 = 24$ 个主要条件。

校准集、测试集和留出集均从 WikiText-2 原始数据集 (`wikitext-2-raw-v1`) 中构建。将原始划分合并、规范化、筛选、去重后，使用随机种子 42 进行确定性打乱；前 512 个样本构成校准池，接下来的 256 个构成测试集，最后 256 个构成留出集。校准仅用于估计激活统计量；每次运行最多使用 32 个校准样本，论文中的指标则基于留出集计算。提供的 Llama-3-8B 评估 JSON 标明了 `data/evaluation.txt`、最多 256 个评估样本、最大序列长度为 256，以及评估的 token 数为 36,153。仅凭该 JSON 无法独立验证评估文件的来源。

全精度模型作为参考模型。对于 Llama-3-8B，报告的检查点大小是序列化叠加文件大小；其中不包含所有未更改的全精度参数。Llama-3-8B 的运行时间测量采用模型分片执行，不应将其视为经过优化的压缩内核性能。

参数指标包括均方误差、均方根误差、平均绝对误差、相对误差、余弦相似度、峰值信噪比和信噪比；功能指标包括在分词和样本匹配的条件下计算的 logit 均方误差/余弦相似度以及下一词元困惑度。存储指标包括选定权重载荷、压缩率、每参数比特数和序列化检查点大小。由于检查点可能是叠加层，因此会将其大小与独立部署存储区分开来。我们还会在可行时记录量化/评估运行时间和 GPU 内存。

L 的评估　　根据评估记录，lama-3-8B 在四个 GPU 上采用模型分片

表 1：主要量化配置。

参数	设置
向量/投影维度 (n, m)	(256，256)
残差阶段 r	1, 2, 3,5
统计精度	32 位
块批大小	4
激活感知缩放	$\beta = 0.5$
变换候选项 K	4
校准样本数 / 序列长度	32 / 256
渐进式 / 熵编码	已启用 / 已启用
随机种子	42

表2：GPT-2 Large 在全部三个量化范围下的比较。全精度困惑度为 29.89。运行时间比为 $T_{\mathrm{FP}}/T_{\mathrm{streaming}}$。

范围	r	参数均方误差	相对误差	Logit 均方误差	Logit 余弦相似度	KL	下一个 (%)	Top-k (%)	PPL	大小 (MiB)	前向传播 (毫秒)	运行时间比
注意力	1	9.225×10^{-4}	0.6100	3.7176	0.9530	2.5918	25.85	30.53	429.36	58.83	731.19	0.0220
注意力	2	3.303×10^{-4}	0.3650	0.6805	0.9919	0.3289	69.11	70.77	43.58	115.97	1295.36	0.0124
注意力	3	1.467×10^{-4}	0.2432	0.2542	0.9969	0.1114	80.56	81.52	34.64	172.95	1776.10	0.0090
注意力	5	5.519×10^{-5}	0.1492	0.0836	0.9990	0.0342	88.86	88.91	31.44	287.32	2909.18	0.0055
MLP	1	8.912×10^{-4}	0.6088	8.2450	0.8936	5.1919	1.35	4.23	5765.20	115.80	1010.37	0.0158
MLP	2	3.192×10^{-4}	0.3644	1.9571	0.9745	1.0087	46.47	52.21	69.61	229.37	1858.88	0.0084
MLP	3	1.418×10^{-4}	0.2429	0.6814	0.9913	0.3180	67.61	70.96	36.94	342.71	2671.84	0.0059
MLP	5	5.348×10^{-5}	0.1491	0.2256	0.9972	0.0952	80.68	82.81	31.15	570.18	4281.04	0.0037
注意力+多层感知机	1	9.016×10^{-4}	0.6092	8.1806	0.8938	5.3807	1.97	7.86	6768.31	174.38	1757.09	0.0091
注意力+多层感知机	2	3.229×10^{-4}	0.3646	2.4858	0.9668	1.4073	40.73	44.64	116.48	345.09	3204.37	0.0050
注意力+多层感知机	3	1.434×10^{-4}	0.2430	0.8028	0.9894	0.3682	65.74	68.23	41.09	515.42	4429.88	0.0036
注意力+多层感知机	5	5.405×10^{-5}	0.1492	0.2542	0.9967	0.1041	80.05	81.55	32.16	857.25	7430.52	0.0022

元数据；GPT-2 Large 的执行设置保持不变。

独立的基线比较方案。额外的联合范围实验（附录 B）采用独立的 FP16 评估，使用 4,096 个输入词元（4,095 个计分词元）、所选权重的估算负载每权重比特数，以及重建的稠密权重。其码率和困惑度测量结果无法与特定范围的流式叠加实验直接比较。

6 结果

我们在残差深度 $r \in \{1、2、3、5\}$ 下评估 GPT-2 Large 的仅注意力、仅多层感知机以及注意力+多层感知机联合量化。评估指标包括重建误差和 logit误差、词元一致率、困惑度、序列化叠加大小以及流式运行时间。

6.1 GPT-2 Large

全精度 GPT-2 Large 在留出数据上的困惑度为 29.89。三种量化范围均采用相同的评估协议，从而能够直接比较不同范围和残差深度下的结果。

6.1.1 不同量化范围比较

图1和表2显示，三个范围均呈现一致的渐进恢复趋势。参数均方误差和相对误差均随 r 以相似方式下降。在 $r = 5$ 时，参数均方误差约为

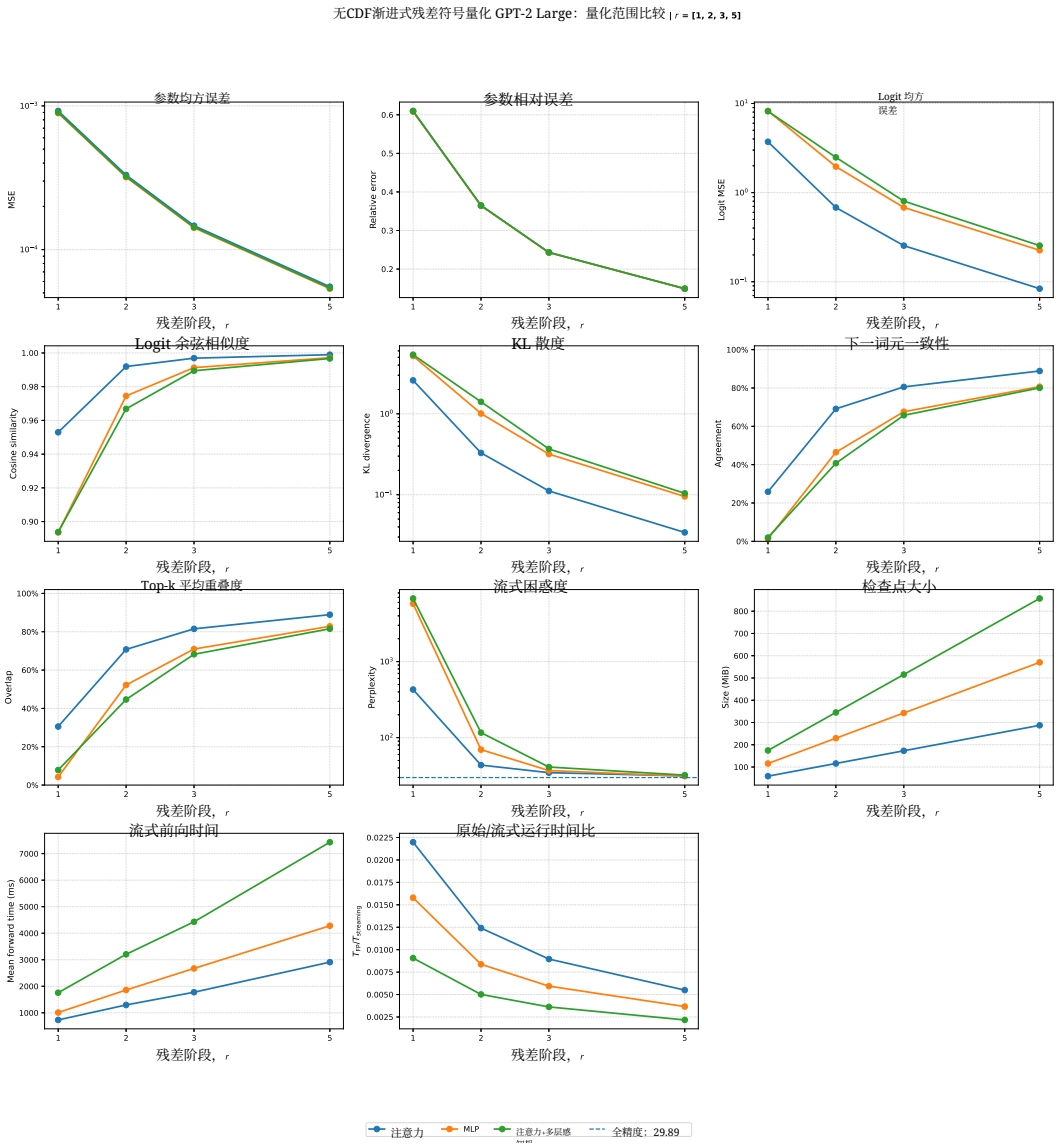

图1：量化　　范围对比：GPT-2 Large 在不同残差深度下的表现 $r \in \{1, 2, 3, 5\}$。

5.3–5.5×10^{-5}，且所有范围内的相对误差均约为 0.149。因此，主要差异体现在模型层面的保真度，而非局部权重重构趋势。

在低和中等残差深度下，仅注意力量化是干扰最小的范围。在 $r = 1$ 时，其困惑度为 429.36，而仅多层感知机量化和联合注意力+多层感知机量化的困惑度分别为 5765.20 和 6768.31。随着残差阶段增加，这一差距迅速缩小：在 $r = 3$ 时，注意力、多层感知机和注意力+多层感知机的困惑度分别为 34.64、36.94 和 41.09。在 $r = 5$ 时，三种配置的表现都接近全精度基线，困惑度分别为 31.44、31.15 和 32.16。

输出保真度也趋于收敛。在 $r = 5$，仅注意力量化展现出最强的整体逻辑值保真度，逻辑值均方误差为 0.0836，余弦相似度为 0.9990，KL 散度为 0.0342，下一词元一致性为 88.86%，top-k 重叠率为 88.91%。仅多层感知机量化和联合量化也恢复了较高的保真度，下一词元一致性分别为 80.68% 和 80.05%。在这一保留的最大阶段数下，仅多层感知机量化取得了最低的困惑度 31.15，而仅注意力

Ta表 3：Llama-3-8B 范围结果（FP 困惑度：11.07；运行时间比：$T_{\text{FP}}/T_{\text{streaming}}$）

范围	r	参数均方误差	相对误差	Logit 均方误差	Logit 余弦相似度	KL	下一个 (%)	前-k (%)	PPL	大小 (MiB)	前向传播 (毫秒)	运行时间比
注意力	1	$3.278e-04$	1.1949	10.9461	0.6640	5.2465	9.62	14.72	2076.92	324.62	1762.50	0.0202
注意力	2	$1.288e-04$	0.7490	4.9474	0.8259	1.0956	55.77	54.29	32.09	645.72	3382.69	0.0105
注意力	3	$5.892e-05$	0.5066	0.9293	0.9543	0.2358	78.24	76.20	13.75	966.73	5030.83	0.0070
注意力	5	$2.247e-05$	0.3129	0.3185	0.9840	0.0789	87.06	85.49	11.90	1610.37	8067.03	0.0044
MLP	1	$5.228e-05$	0.6118	12.0770	0.1970	10.3045	0.05	0.66	327947.62	1350.58	18748.79	0.0019
MLP	2	$1.875e-05$	0.3664	4.2837	0.8565	1.2143	54.55	53.25	35.70	2696.00	24639.92	0.0014
MLP	3	$8.345e-06$	0.2444	1.4481	0.9382	0.4110	72.71	70.70	16.04	4040.53	31273.26	0.0011
MLP	5	$3.152e-06$	0.1502	0.5938	0.9738	0.1579	83.11	81.06	12.55	6737.66	43173.52	0.0008
注意力+多层感知机	1	$1.053e-04$	0.8189	14.3613	0.0109	11.6030	0.00	0.42	1211639.22	1674.99	21097.99	0.0017
注意力+多层感知机	2	$3.991e-05$	0.5043	8.5083	0.7659	2.4650	36.70	37.13	123.44	3341.50	28899.10	0.0012
注意力+多层感知机	3	$1.807e-05$	0.3393	2.0122	0.9171	0.4889	69.33	67.32	17.29	5007.07	37183.52	0.0009
注意力+多层感知机	5	$6.867e-06$	0.2092	0.6937	0.9692	0.1704	81.61	79.64	12.73	8347.85	50059.68	0.0007

量化保留了更强的词元级和 logit级一致性。

存储和运行时间主要随量化参数的数量而变化。仅注意力量化约 2.36 亿个参数，仅多层感知机量化约 4.72 亿个参数，联合范围则量化约 7.08 亿个参数。因此，在 $r = 5$ 下，检查点大小从注意力的 287.32 MiB 增至多层感知机的 570.18 MiB，以及注意力+多层感知机的 857.25 MiB。流式前向时间也呈现相同的递增顺序：约 2.91 秒、4.28 秒和 7.43 秒。这些结果表明，残差深度决定范围内的保真度，而范围大小则决定主要的存储和运行时间成本。

总体而言，$r = 5$（本研究中保留的最大残差深度）使三个范围的困惑度均接近基线。因此，结果支持一种清晰的率失真解读：在所评估的范围内增加残差深度能够恢复模型质量，而量化更大的子系统，则需要付出更高成本才能达到相近的保真度水平。

6.2 Llama-3-8B

全精度 Llama-3-8B 模型（`meta-llama/Meta-Llama-3-8B`）在提供的评估集（36,153 个计分词元）上，困惑度为 11。07。我们在 $r \in \{1, 2, 3, 5\}$ 下评估仅注意力量化、仅多层感知机量化，以及注意力+多层感知机联合量化。这些范围分别包含 1,342,177,280、5,637,144,576 和 6,979,321,856 个选定参数。这些是特定范围的量化实验，并非全模型独立压缩率。

6.2.1 不同量化范围比较

图2 和表3 展示了各个范围内的渐进式恢复。在 $r = 1$时，注意力的困惑度为 2076。92，多层感知机为 327947。62，联合量化为 1.21×10^6 。因此，在当前设置下，单个残差阶段无法保留可用的语言模型保真度，尤其是在量化多层感知机权重时。下一词元一致率分别为 9。62%、0。05% 和 0。00%。

在 $r = 2$ 时，注意力、多层感知机和联合范围的困惑度分别降至 32.09、35.70 和 123.44。在 $r = 3$ 时，分别达到 13.75 、 16.04 和 17.29。采用五个残差阶段时，相应的困惑度分别为 11.90 、12.55 和 12.73，而全精度下为 11.07。在 $r = 5$ 时，相对增幅分别约为 7.5%、13.3% 和 15.0%。

输出保真度随着困惑度的降低而提升。在 $r = 5$ 下，仅注意力设置达到 Logit 余弦相似度 0.9840、KL 散度 0.0789、下一词元一致性 87.06%，以及 top-k 重叠率 85.49%。仅多层感知机设置的余弦相似度为 0.9738，一致率为 83.11%，top-k 重叠率为 81.06%。联合量化的余弦相似度为 0.9692，一致率为 81.61%，top-k 重叠率为 79.64%。在每个范围内，参数均方误差会随着阶段深度增加而降低：从 3.278×10^{-4} 到 2.247×10^{-5}

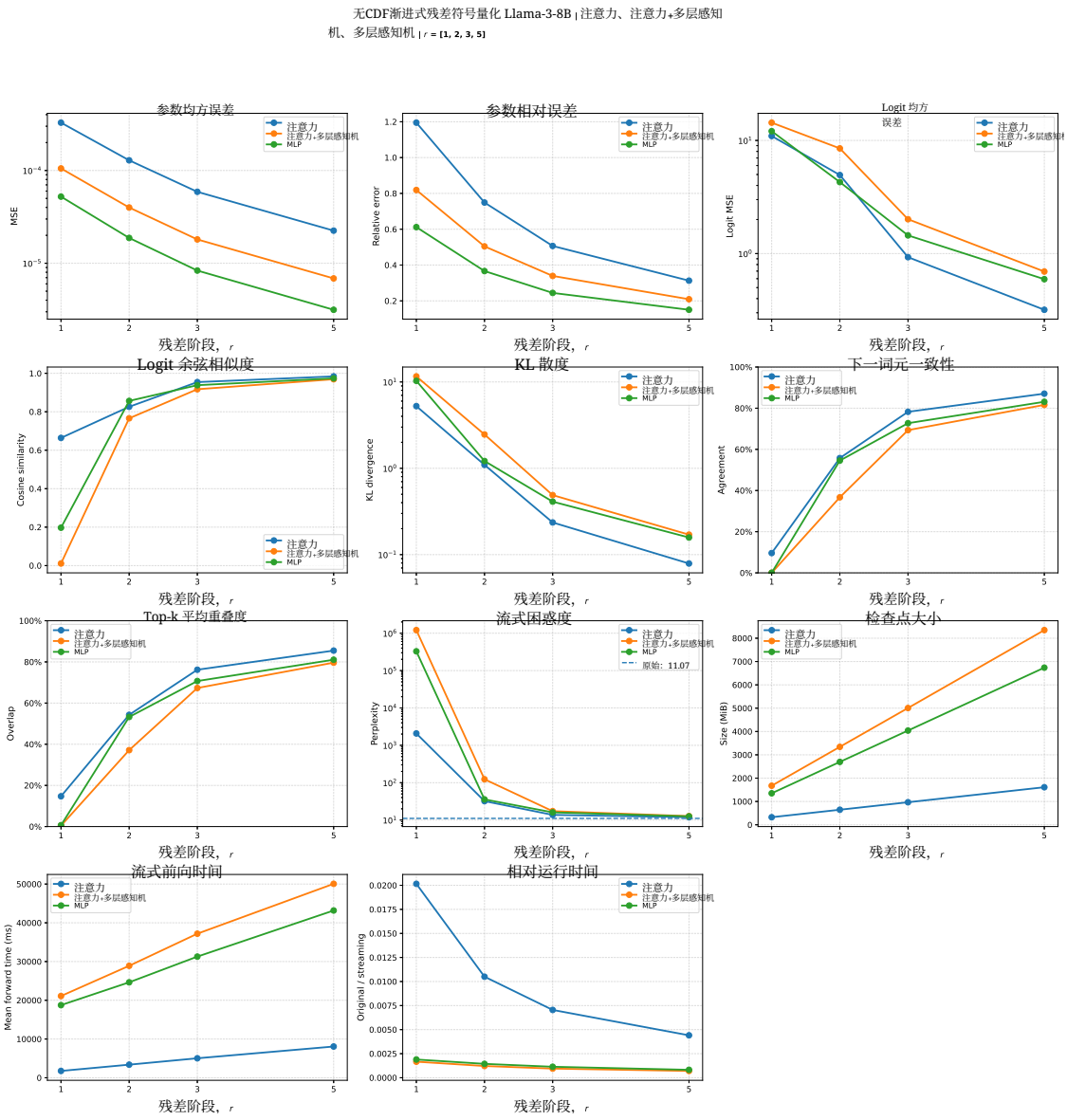

图2：Llama-3-8B 在不同残差深度下的量化范围比较 $r \in \{1,2,3,5\}$。

注意力为 5.228×10^{-5} 到 3.152×10^{-6}，多层感知机为 1.053×10^{-4} 到 6.867×10^{-6}，联合量化为。

在 $r = 5$，对于仅注意力、仅多层感知机和联合范围，序列化叠加检查点分别占用 1610.37 MiB、6737.66 MiB 和 8347.85 MiB。平均流式前向时间分别为 8.07 秒、43.17 秒和 50.06 秒。流式路径明显慢于全精度参考（每次前向传播约 35 毫秒）；这些测量结果并未证明推理速度得到加速。这些检查点还保留了对未更改基础模型权重的依赖，因此不应被理解为自包含的部署产物。

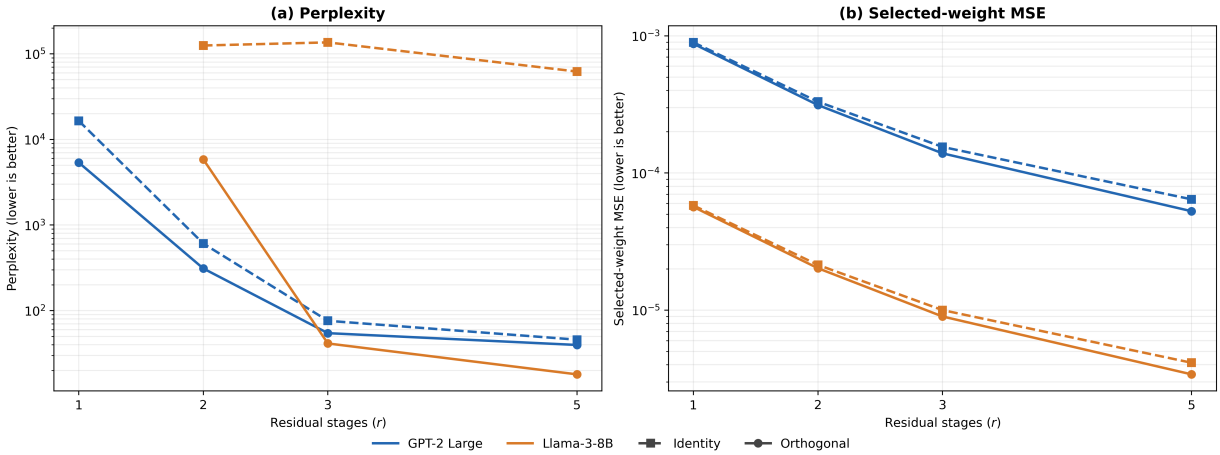

图 3：GPT-2 Large 和 Llama-3-8B 的正交混合与模型保真度。

6.3 跨模型分析

GPT-2 Large 和 Llama-3-8B 呈现出相同的总体趋势：增加残差深度会降低参数失真，并提升仅注意力、仅多层感知机和联合量化的语言模型保真度。然而，它们的基线困惑度和量化参数数量不同，因此不应将绝对困惑度和原始检查点大小理解为可直接对应的比较结果。GPT-2 Large 的全精度困惑度为 29.89，而 Llama-3-8B 为 11.07。

在 $r = 5$，GPT-2 Large 在注意力、多层感知机和联合范围下的困惑度分别为 31.44、31.15 和 32.16，较基线分别高出约 5.2%、4.2% 和 7.6%。Llama-3-8B 对应的困惑度分别为 11.90、12.55 和 12.73，较基线分别高出 7.5%、13.3% 和 15.0%。对于这两个模型，在 $r = 1$ 时，仅注意力量化的影响最小；但在 $r = 5$ 时，各范围之间的差异仍然明显，Llama-3-8B 尤其如此。

Llama-3-8B 较大的量化范围还会产生大得多的序列化叠加层，并延长流式重建时间。这些数据反映的是经测量的原型，而非经归一化的压缩优势或硬件加速。结果表明，残差深度可作为一种有效的保真度调节手段，同时也凸显了进行范围感知和实现感知评估的必要性。

6.4 正交混合比较

一项独立的联合范围基准测试在 4,095 个计分词元上评估 FP16 重建的稠密权重；其结果无法直接与特定范围的流式处理运行结果进行比较。图3显示，在所有测试深度下，正交混合都能降低 GPT-2 Large 的困惑度和选定权重均方误差。在 Llama-3-8B 上，每个深度的均方误差均有所改善，且困惑度在 $r = 2$、3、5 处有所改善；省略的 $r = 1$困惑度在采用正交混合时更差（2.42×10^6 而非 9.03×10^5）。两种符号变体的估计载荷几乎相同。

标量和向量基线。 正交符号编解码器在每个选定权重载荷对应的困惑度方面并不优于传统量化器。在 $r = 5$，GPT-2 Large 符号编码以 6.618 的估计每权重比特数取得 39.73 的困惑度，而 $r = 1$ 的标量 4 位编码以 3.658 每权重比特数取得 37.59 的困惑度，残差向量量化（$K = 256$，$r = 5$）则以 4.995 每权重比特数取得 36.98 的困惑度。对于 Llama-3-8B，符号编码以 6.611 每权重比特数取得 18.02 的困惑度，相比之下，标量 4 位编码以 3.632 每权重比特数取得 16.05 的困惑度；经测试的向量量化（$K = 256$、

$r = 5$）则以 4.948 每权重比特数取得 8449 的困惑度。因此，该基准测试支持正交混合和渐进式符号细化，而非表明其具有普遍更优的率失真前沿。完整的比较结果及测量注意事项见附录 B。

7 消融研究

除非另有说明，两个模型均采用联合注意力+多层感知机量化，设置为 $r = 5$，使用 $n = m = 256$、32 位统计信息、128 个块行、块批大小为 4、 $K = 4$、激活感知缩放（AAS）、32 个校准样本、单调阶段接受以及随机种子 42。详细测量结果见附录 A。

对于 GPT-2 Large，AAS 将 Logit 均方误差从 0.3001 降至 0.2543，并将困惑度增幅从 9.79 % 降至 7.63%，尽管参数均方误差更高。四个变换候选项优于一个候选项；校准大小和块行效应较为有限。与使用 32 位统计信息相比，使用 16位统计信息会将压缩率从 $4.77\times$提高至 $5.53\times$，而保真度几乎没有变化。向量维度决定了观测到的存储–保真度权衡。

对于 Llama-3-8B，AAS 将 Logit 均方误差从 1.0983 降至 0.6920，并将 PPL 从 13.9445 降至 12.5854，尽管参数均方误差从 3.379×10^{-6} 升至 6.867×10^{-6}。16位统计信息会将叠加层从 8347.85 MiB 降至 7220.71 MiB，同时 PPL 相近。将 K 从 1 增加到 4 后，PPL 从 12.6102 降至 12.5854，但编码时间从约 100.3 秒增加到 175.6 秒。其他测试设置的影响较小，或取决于模型。没有可用的 Llama-3-8B 向量维度或提前停止率消融结果。这些结果表明，应将功能保真度与权重误差结合起来进行评估。

8 讨论

编码、深度与范围。 我们提出的框架结合了按失真选择的正交混合、解析符号重建和渐进式残差细化，无需学习型码本或 CDF 估计。增加 r 会以存储换取保真度：从 $r = 1$ 到 $r = 5$，联合范围困惑度从 6768.31 降至 32.16（GPT-2 Large），从 1.21×10^6 降至 12.73（Llama-3-8B），而全精度结果分别为 29.89 和 11.07。在 $r = 1$下，仅注意力编码的干扰较小；在 $r = 5$ 时，所有范围都接近基线（第 6 节）。

权重误差与输出保真度。 权重均方误差并不能可靠地预测功能质量。在$r = 5$下，Llama-3-8B 的仅注意力编码和联合编码的均方误差分别为 2.247×10^{-5} 和 6.867×10^{-6}，但困惑度分别为 11.90 和 12.73；两者的均方误差覆盖了不同的参数集。GPT-2 Large 的消融实验同样显示，尽管权重误差更高，AAS 仍能带来收益。在 Llama-3-8B 上，AAS 将权重均方误差从 3.379×10^{-6} 提高至 6.867×10^{-6}，但将 Logit 均方误差从 1.0983 降至 0.6920，并将困惑度从 13.9445 降至 12.5854。

校准与工作点。 两个模型的性能都随深度增加而提升，但对低深度的敏感程度不同。Llama-3-8B 使用 AAS $\beta = 0.5$ 和四种变换候选项。单因素消融涵盖 AAS、候选项数量和校准数量、统计精度、块行数及批量大小，但未考察交互作用、向量维度或随机种子泛化。较低的 r 和较大的向量有利于存储；较高的 r 和较小的向量有利于保真度。对于两个模型， $r = 3$ 都能显著恢复质量， $r = 5$ 则接近基线；16 位统计量可进一步减少存储开销。增加深度或范围会增加存储开销和重建时间。流式处理仍慢于全精度，而叠加层大小未计入独立部署成本。因此，跨范围速率需要以选定权重每字节位数或压缩率来衡量。

基线与高斯动机。 附录 B 比较了恒等/正交符号编码、标量量化和残差向量量化在估计选定权重速率下的表现。混合通常能改善符号编码，但标量量化或向量量化可能以更低的每权重比特数实现更低的困惑度；无码本编码既不保证推理加速，也不保证在所有情况下都具有更优的率失真表现。最后，中心极限定理为混合提供了动机，但并不要求假设高斯性：重建使用的是最小二乘拟合增益，而非高斯累积分布函数计算。

9 局限性

AAS 依赖有限的校准数据，域偏移鲁棒性尚未经过测试。全局残差深度忽略了层特定敏感性。不同量化范围的参数数量不同，因此若不进行归一化，叠加层大小便无法比较，且与独立部署存储有所区别。流式权重重建比全精度处理更慢，而多 GPU 扩展则取决于硬件和传输开销。

评估仅涉及 GPT-2 Large 和 Llama-3-8B，主要采用困惑度和短时域保真度指标；下游任务、长上下文、指令遵循和生成鲁棒性均尚未测试。尽管 Llama-3-8B 的评估包含范围/深度扫描和组件消融，但尚未进行向量维度、提前停止、重复随机种子和匹配速率基线分析。单因素消融未考虑超参数交互，广泛的硬件匹配训练后量化比较也仍然缺失。最后，高斯中心极限定理行为只是研究动机，并非经验证的结果；正交性本身既不能保证高斯性，也不能保证独立性。

10 结论与未来工作

我们提出了一种免训练、无 CDF、受高斯分布-中心极限定理启发的残差符号量化方法，结合正交混合与紧凑阶段统计量。渐进式重建遵循

$$\boxed{\hat{x}^{(r)} = \sum_{k=1}^{r} \hat{e}^{(k)}},$$

其中，残差深度 r 控制率失真权衡，无需学习型码本、经验累积分布函数或网格优化。

渐进式细化提升了仅注意力、仅多层感知机和联合量化的保真度。在 $r = 5$下，GPT-2 Large 对应的困惑度分别为 31.44、31.15 和 32.16（全精度：29.89）；Llama-3-8B 对应的困惑度分别为 11.90、12.55 和 12.73（全精度：11.07），保真度差距取决于范围。

Llama-3-8B 消融实验（第 7 节；附录 A）表明，激活感知缩放将困惑度从 13.9445 降至 12.5854，并将下一词元一致性从 75.68% 提升至 81.75%，尽管参数均方误差有所增加。将统计精度从 32 位降至 16 位，可将检查点大小从 8347.85 MiB 缩小至 7220.71 MiB，对应的困惑度分别为 12.5854 和 12.5803。与一个变换候选项相比，四个变换候选项能小幅改善均方误差和困惑度，但编码成本更高；在测试范围内，校准规模和块行/批次大小对困惑度的影响极小。这些消融实验未涵盖正交与恒等比较及自适应提前停止。

另一项匹配配置联合范围基准测试（附录 B）表明，在 $r = 5$ 时，与估计选定权重相近的恒等符号编码相比，正交混合使 GPT-2 Large 的困惑度从 45.81 降至 39.73，并使 Llama-3-8B 的困惑度从约6.22×10^4 降至 18.02

载荷。不过，在一些较高载荷率下，标量量化可实现更低的困惑度，因此不能声称其在率失真方面始终占优。

流式重建能够保持紧凑性和保真度，但会带来显著的译码与前向传播开销。未来的工作包括由敏感度或残差能量引导的、特定于层和投影的阶段分配，

$$\boxed{\min_{\{r_\ell\}} D_{\text{functional}} \quad \text{s.t.} \quad \sum_{\ell=1}^{L} S_\ell(r_\ell) \le B,}$$

融合译码／矩阵乘法内核、改进的变换选择、自包含的压缩检查点、更广泛的下游任务和长上下文评估、更多模型系列，以及匹配的训练后量化基线。

参考文献

Ashkboos, S., Mohtashami, A., Croci, M. L., Li, B., Cameron, P., Jaggi, M., Alistarh, D., Hoefler, T. 和 Hensman, J. Quarot：旋转大型语言模型中的无离群值 4位推理。载于 Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. 和 Zhang, C.（编），神经信息处理系统进展，第 37 卷，第 100213–100240 页。Curran Associates, Inc.，2024年。doi：10.52202/079017-3180。

Dettmers, T., Lewis, M., Belkada, Y. 和 Zettlemoyer, L. LLM.int8()：面向大规模 Transformer模型的8位矩阵乘法。载于神经信息处理系统进展，第 35 卷，第 30318–30332 页，2022年。

Dettmers, T., Pagnoni, A., Holtzman, A. 和 Zettlemoyer, L. QLoRA：量化大型语言模型的高效微调。神经信息处理系统进展，36:10088–10115，2023。

Egiazarian, V., Panferov, A., Kuznedelev, D., Frantar, E., Babenko, A. 和 Alistarh, D. 通过加性量化实现大语言模型的极致压缩。载于第*41*届国际机器学习会议论文集，机器学习研究论文集第 235 卷，第 12284–12303 页。PMLR，2024年。

Frantar, E., Ashkboos, S., Hoefler, T. 和 Alistarh, D. GPTQ：面向生成式预训练Transformer的精准训练后量化。发表于国际学习表征会议，2023。

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. 和 Chen, W. LoRA：大语言模型的低秩适配。发表于国际学习表征会议，2022年。

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C. 和 Han, S. AWQ：面向端侧大型语言模型压缩与加速的激活感知权重量化。载于 Gibbons, P., Pekhimenko, G. 和 Sa, C. D.（编），机器学习与系统论文集，第 6 卷，第 87–100 页，2024年。

Liu, Z.、Zhao, C.、Fedorov, I.、Soran, B.、Choudhary, D.、Krishnamoorthi, R.、Chandra, V.、Tian, Y. 和 Blankevoort, T. SpinQuant：采用学习到的旋转的大语言模型量化。载于 国际学习表征会议，2025年。

Luo, Y.、Dong, B.、Cheng, W. 和 Shen, H. 循环残差量化：面向大语言模型的渐进式多精度表示，2026。URL `https://arxiv.org/abs/2608.04048`。

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 等. 语言模型是无监督多任务学习器。*OpenAI* 博客，1(8):9，2019年。

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., 等. LLaMA：开放且高效的基础语言模型。*arXiv* 预印本 *arXiv:2302.13971*，2023a。Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., 等. Llama 2：开放的基础模型和微调聊天模型。*arXiv* 预印本 *arXiv:2307.09288*，2023b。Tseng, A., Chee, J., Sun, Q., Kuleshov, V., 和 De Sa, C. QuIP#：通过哈达玛非相干性和格码本实现更优的大语言模型量化。载于第*41*届国际机器学习会议，第235卷机器学习研究论文集，第48630–48656页。PMLR，2024a。Tseng, A., Sun, Q., Hou, D., 和 De, C. Qtip：基于格栅和非相干性处理的量化。载于 Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., 和 Zhang, C.（编），神经信息处理系统进展，第37卷，第59597–59620页。Curran Associates, Inc.，2024b。doi: 10.52202/079017-1904。van Baalen, M., Kuzmin, A., Nagel, M., Couperus, P., Bastoul, C., Mahurin, E., Blankevoort, T., 和 What- mough, P. GPTVQ：维数的祝福助力大语言模型量化。*ar Xiv* 预印本 *arXiv:2402.15319*，2024年。Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., 和 Polosukhin, I. 注意力就是一切。载于 Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., 和 Garnett, R.（编），神经信息处理系统进展，第30卷。Curran Associates, Inc.，2017。Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., 和 Han, S. SmoothQuant：面向大语言模型的精准高效训练后量化。载于国际机器学习会议，第38087–38099页。PMLR，2023。Zandieh, A., Daliri, M., Hadian, M., 和 Mirrokni, V. TurboQuant：具有近乎最优失真率的在线向量量化。载于国际学习表征会议，2026。

详细消融研究

本附录提供了第 7 节中总结的完整组件级消融实验。GPT-2 Large 和 Llama-3-8B 的主要结果，包括其范围比较表、图表和详细解读，均已在第 6 节中呈现；此处不再重复。

除非另有说明，两个模型均采用联合注意力+多层感知机量化，设置为 $r = 5$, $n = m$ =256、32 位统计信息、128 个块行、块批大小 4、 $K = 4$、激活感知缩放（AAS）、32 个校准样本、单调阶段接受和随机种子 42。仅改变指定组件。

对于 GPT-2 Large（表 4），尽管参数均方误差更高，AAS 仍提升了功能保真度：Logit 均方误差下降了 $0.3001 \rightarrow 0.2543$，PPL 下降了 $\Delta 9.79\% \rightarrow 7.63\%$。 $K = 4$ 优于 $K = 1$；校准规模的影响较小。与 32 位统计量相比，使用 16位统计量可提高压缩率 $4.77\times \rightarrow 5.53\times$，且质量变化可忽略不计。块行变化较小，而向量维度则呈现出预期的保真度与压缩率权衡。

表 4：GPT-2 Large 组件消融（$r = 5$，联合范围）。

消融	设置	参数均方误差↓	Logit 均方误差↓	△困惑度（%）↓	每权重比特数↓	压缩率↑
AAS	Off	5.204×10^{-5}	0.3001	9.79	6.71	4.77×
	On	5.405×10^{-5}	0.2543	7.63	6.71	4.77×
变换	$K = 1$	5.471×10^{-5}	0.2684	8.36	6.71	4.77×
	$K = 4$	5.405×10^{-5}	0.2543	7.63	6.71	4.77×
校准	4	5.548×10^{-5}	0.2579	7.47	6.71	4.77×
	8	5.462×10^{-5}	0.2566	7.42	6.71	4.77×
	16	5.423×10^{-5}	0.2546	7.93	6.71	4.77×
	32	5.405×10^{-5}	0.2543	7.63	6.71	4.77×
统计量	16位	5.405×10^{-5}	0.2537	7.63	5.79	5.53×
	32位	5.405×10^{-5}	0.2543	7.63	6.71	4.77×
块行	64	5.380×10^{-5}	0.2548	7.55	6.74	4.74×
	128	5.405×10^{-5}	0.2543	7.63	6.71	4.77×
	256	5.422×10^{-5}	0.2586	7.71	6.69	4.78×
向量 (n， m)	(128，128)	5.176×10^{-5}	0.2449	6.92	8.40	3.81×
	(256， 256)	5.405×10^{-5}	0.2543	7.63	6.71	4.77×
	(512， 512)	5.526×10^{-5}	0.2622	8.43	5.87	5.45×

表 5：Llama-3-8B 组件消融实验（$r = 5$，联合范围）。

消融	设置	参数均方误差↓	Logit 均方误差↓	Δ困惑度（%）↓	每权重比特数↓	压缩率↑
AAS	Off	3.379×10^{-6}	1.0983	26.10	6.65	4.81×
	On	6.867×10^{-6}	0.6920	13.81	6.66	4.81×
变换	$K = 1$	6.944×10^{-6}	0.6994	14.04	6.66	4.81×
	$K = 4$	6.867×10^{-6}	0.6920	13.81	6.66	4.81×
校准	4	—	0.6804	14.45	6.66	4.81×
	8	—	0.6561	13.55	6.66	4.81×
	16	—	0.6769	13.80	6.66	4.81×
	32	—	0.6920	13.81	6.66	4.81×
统计量	16位	—	0.6904	13.77	5.76	5.55×
	32位	—	0.6920	13.81	6.66	4.81×
块行	64	—	0.6730	14.00	6.70	4.77×
	128	—	0.6920	13.81	6.66	4.81×
	256	—	0.6447	14.48	6.63	4.83×
块批次	1	—	0.6918	13.84	6.66	4.81×
	4	—	0.6920	13.81	6.66	4.81×
	8	—	0.6918	13.84	6.66	4.81×

注意：破折号表示所提供的表格中没有足够精度的参数均方误差值；它们不代表零。

本文　表格报告了所有经过测试的消融实验；主要发现总结于第 7 节。

B 额外联合范围量化器基准

此基准比较 GPT-2 Large 和 Llama-3-8B 的联合注意力+多层感知机权重上的标量、残差向量、恒等符号和正交符号量化器。它使用 FP16 重建的稠密权重、256 维向量/块维度、32位符号统计量、4,096 个词元的输入（4,095 个计分词元），以及 $r \in \{1、2、3、5\}$。标量基线采用 2 位或 4 位残差量化；残差向量量化采用 8 维向量、大小为 16 或 256 的码本，以及五次迭代。与范围扫描不同，此基准评估重建的稠密权重，因此其困惑度不能直接比较。

有效载荷和运行时间。 估算的每权重比特数（BPW）等于估算有效载荷字节数乘以八，再除以选定参数的数量。此数值不包括未量化权重、检查点封装和运行时存储。稠密权重评估无法证明压缩推理加速；编码器耗时取决于具体实现。比较有效载荷，而不是相同的残差深度。

渐进式细化与混合。 在图 4 中，GPT-2 Large 的正交符号编码将困惑度从 5368。89 和均方误差 8.755×10^{-4}（$r = 1$，1.319 每权重比特数）提升至困惑度 39。73 和均方误差 5.265×10^{-5}（$r = 5$，6.618 每权重比特数）。$r = 5$ 处的恒等符号编码达到困惑度 45。81 和均方误差 6.428×10^{-5}（6.611 每权重比特数）。在图 5 中，Llama-3-8B 的正交符号编码将困惑度从 2.425×10^{6} 和均方误差 5.660×10^{-5}（$r = 1$，1.313 每权重比特数）提升至困惑度 18。02 和均方误差 3.405×10^{-6}（$r = 5$，6.611 每权重比特数）；$r = 5$ 处的恒等符号编码得到困惑度 62178。91 和均方误差 4.136×10^{-6}（6.602 每权重比特数）。然而，在 $r = 1$ 处，正交符号编码的 Llama-3-8B 困惑度劣于恒等编码（2.425×10^{6} 对比 9.03×10^{5}）。

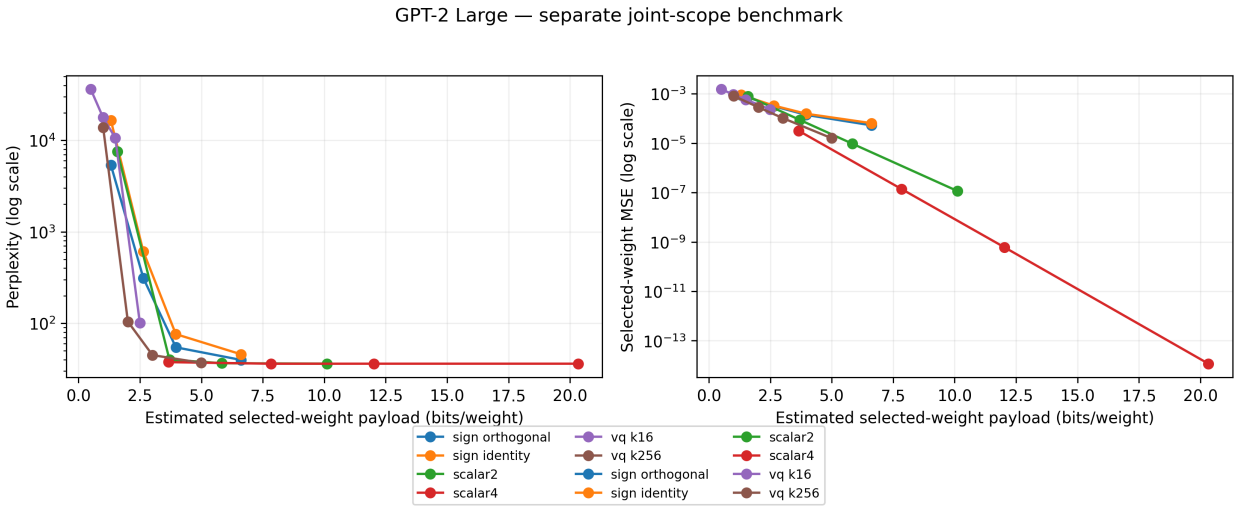

图 4：GPT-2 Large 联合范围量化器对比。

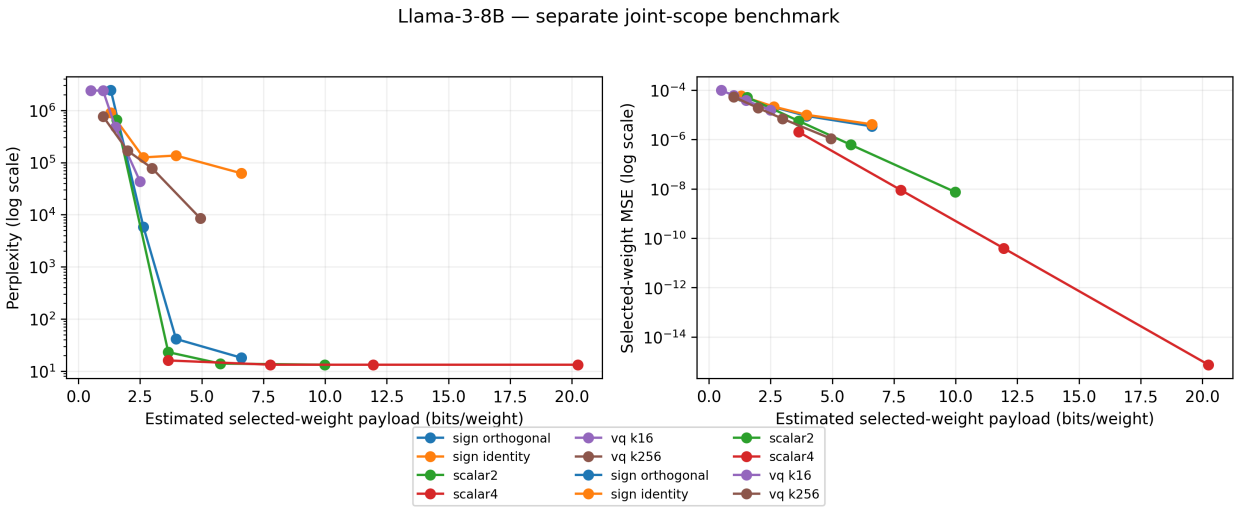

图 5：Llama-3-8B 联合范围量化器对比。

标量和向量基线。 在 GPT-2 Large 上， $r = 1$ 处的 4 位标量达到 37.59 的困惑度，码率为 3.658 每权重比特数；残差向量量化（$K = 256, r = 5$）达到 36.98 的困惑度，码率为 4.995 每权重比特数。在这一速率–困惑度比较中，两者均优于 $r = 5$ 处的正交符号编码（39.73，6.618 每权重比特数）。在 Llama-3-8B 上， $r = 1$ 处的 4 位标量达到 16.05 的困惑度，码率为 3.632 每权重比特数；相比之下，正交符号编码的困惑度为 18.02，码率为 6.611 每权重比特数。残差向量量化（$K = 256, r = 5$）的困惑度为 8449，码率为 4.948 每权重比特数；其表现远差于符号编解码器，但码率更低。符号编码器的速度快于受测的向量量化实现，但计时结果未按码率或计算内核进行匹配。1 位标量条目的有效残差阶段数为零，估算负载可忽略不计，困惑度则极其糟糕；因此它们并非可用的 1 位基线，也未纳入对比图。这些结果支持渐进式无码本编码和正交混合，但并不能证明其在率失真或推理速度方面普遍占优。