大模型架构创新与底层细节

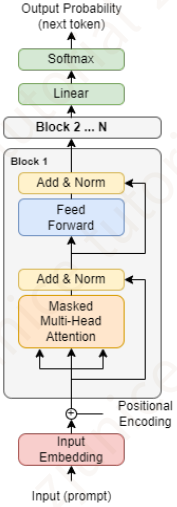
Networked Intelligence for Comprehensive Efficiency (NICE) Lab
College of Information Science and Electronic Engineering
Zhejiang University
http://nice.rongpeng.info/

Aug. 12, 2025

目录

研究背景与问题引入

Decoder-only Transformer：结构与训练/推理



- **块结构**：对 $X \in \mathbb{R}^{L\times d}$ 逐层执行

$$Q=XW_Q,\ K=XW_K,\ V=XW_V,$$

$$A=\mathrm{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}+M_{\text{causal}}\Big)V,$$

$$H=\mathrm{RMSNorm}(X+A),\quad Y=\mathrm{RMSNorm}(H+\mathrm{FFN}(H)).$$

仅**自注意力**并施加**因果掩码** M_{causal}，保持自回归。

- **训练目标**：自回归似然 $\mathcal{L}=-\sum_{t=1}^{L}\log p(x_t\mid x_{<t})$，其中 $p(x_t|x_{<t})=\mathrm{softmax}(W_o h_t)$；梯度只来自**历史上下文**。

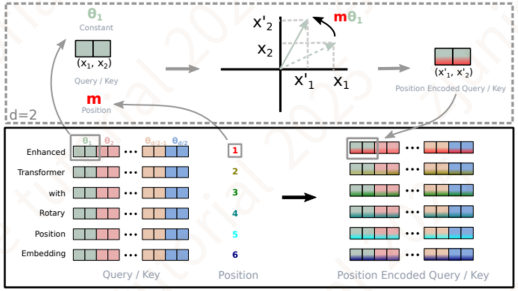
RoPE（旋转位置编码）

- 把 q, k 的**相邻维度成对**视作 2D 向量，按位置 t 旋转角度 $t\theta_i$（每对维度有自己的频率 θ_i）。
- **关键公式**（旋转与相对性）：

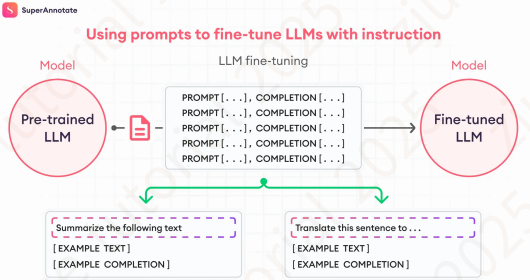
$$\mathrm{Rot}_t \begin{bmatrix} x_{2i} \\ x_{2i+1} \end{bmatrix} = \begin{bmatrix} \cos(t\theta_i) & -\sin(t\theta_i) \\ \sin(t\theta_i) & \cos(t\theta_i) \end{bmatrix} \begin{bmatrix} x_{2i} \\ x_{2i+1} \end{bmatrix}, \quad \theta_i = b^{-2i/d_k}$$

$$\langle \mathrm{Rot}_t(q), \mathrm{Rot}_s(k) \rangle = \langle q, \mathrm{Rot}_{t-s}(k) \rangle$$

- 在**不改注意力公式、几乎不增参数/算子**的前提下，让打分只依赖**相对位置差** $(t-s)$。



监督微调（SFT）

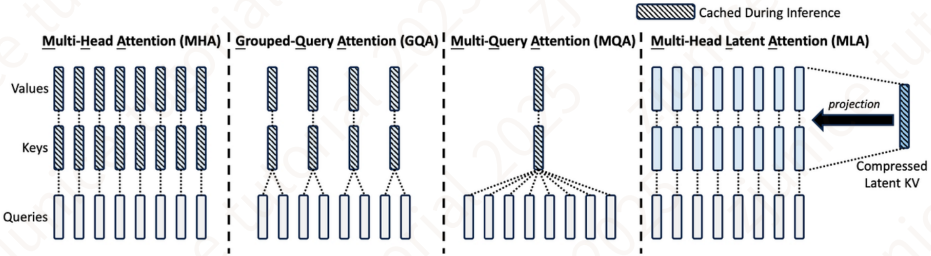


- 在已预训练的 LLM 上，用成对的 指令/上下文 → 目标回答样本做教师强制训练，让模型学会**遵循指令与格式**。
- 全参微调（效果强、资源高）或 LoRA/QLoRA（便宜、易部署）；LoRA 需选择注入模块（Q/K/V/O/FFN 等）与合适 rank。
- 快速**注入任务知识/格式/风格**、打基础；若要**偏好对齐/安全边界**，在 SFT 之后再加 DPO/RLHF 等。
- **常见坑**：数据重复与评测污染、模板过拟合、长度分布极端、只覆盖“简单样本”导致上线后脆弱。

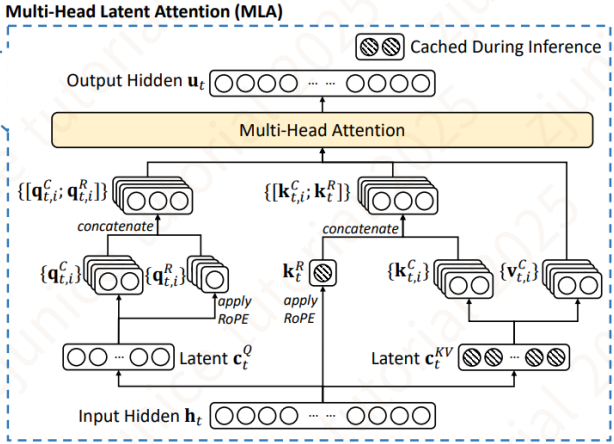
KV-Cache 与注意力压缩

DeepSeek V3：多头潜变量注意力（MLA）

- **动机**：长上下文时，KV 缓存占用极大显存；MLA 将 K, V 压缩进**低维潜空间缓存**，减小显存占用。
- **做法**：编码时 $z_K, z_V = f_{\downarrow}(K, V)$ 并缓存；解码时 $K', V' = f_{\uparrow}(z_K, z_V)$ 再与原 Q 做注意力（接口不变）。
- **收益/成本**：显著降低缓存显存与，**对超长序列尤明显**；但需额外投影/还原算子与校准训练，工程门槛高于 GQA。



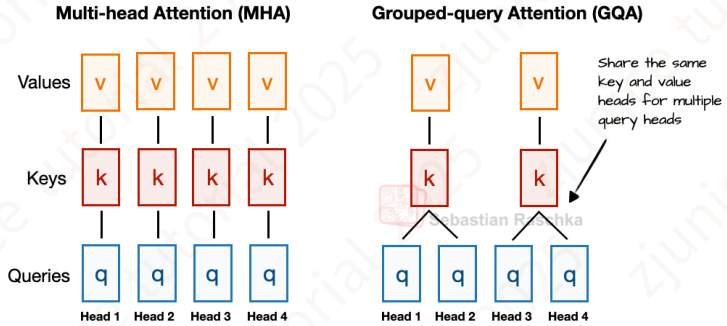
MLA 原理与计算流程



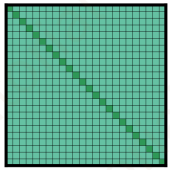
- **编码阶段**：$z_K, z_V = f_{\downarrow}(K, V)$；仅缓存 z_*，不缓存原始 K, V。
- **解码阶段**：$K', V' = f_{\uparrow}(z_K, z_V)$ 与 Q 计算注意力；**不改**注意力算子接口，便于嵌入现有推理栈。
- **关键超参**：潜空间维度 r、上下投影是否共享、是否跨层复用；r 太小会**伤远程依赖**，过大则**内存收益变小**。

Llama 4 GQA 架构创新与选型原因

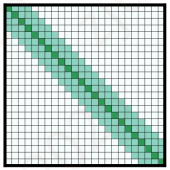
- **机制**：将 H 个头按 G 组共享 K/V；KV 缓存规模近似降为原来的 H/G。
- **取舍**：几乎**零侵入**（沿用标准注意力内核），训练/推理链路兼容；但跨头差异被弱化，少量任务可能有轻微精度损失。
- **工程侧**：对分块 KV、流式解码、张量并行**天然友好**，大批量与短上下文也能稳定获益。



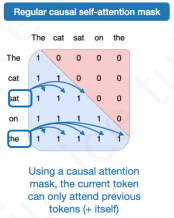
Gemma 3/3n：滑动窗口注意力机制 (SWA)

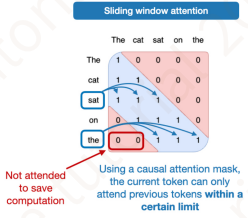


(a) Full n^2 attention



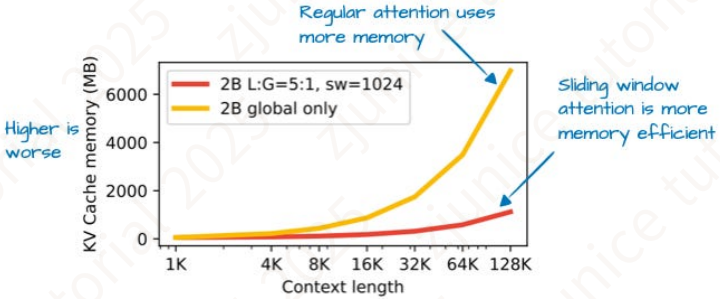
(b) Sliding window attention





- **复杂度**：将感受野限制为窗口 W，单层计算/缓存从 $O(L^2)$ 降到 $O(L \cdot W)$，长序列吞吐**线性增益**。
- **表达力补偿**：仅局部会损失远程依赖；采用“**滑窗层: 全局层** =5:1”，用稀疏全局层**周期性聚合**长距信息。
- **关键值**：窗口 W、全局层周期 k、跨层错位/膨胀窗口等，可缓解边界效应与信息截断。

Gemma 3/3n：滑动窗口注意力 (SWA) 的实现



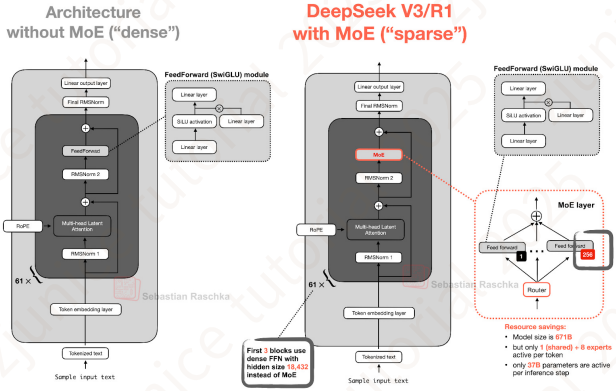
- **KV 缓存优化**：显存占用下降 80-90%，推理门槛极大降低
- **推理速度提升**：长文本吞吐量提升 3-5 倍，支持大批量和端侧部署
- **精度损失极低**：绝大多数任务与全局注意力持平
- **端侧适配**：Gemma 3n 结合滑窗与参数切片，低功耗设备可跑超大模型

总结：KV-Cache 与注意力压缩

模型	机制	上线/开源时间	KV 缓存占用	上下文/吞吐	工程复杂度
DeepSeek V3	MLA	2024-12-27	低（降维缓存）	长/高	较高
LLaMA 4	GQA	2025-04-05	中（KV 共享）	中 长/高	低
Gemma 3	SWA	2025-03-12	低（滑窗）	很长/高	低 中

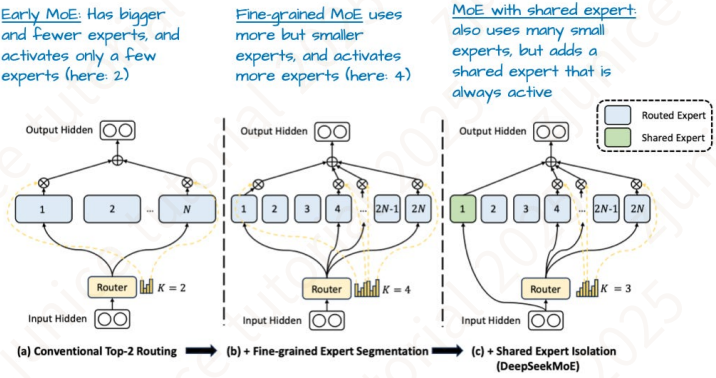
稀疏激活与专家路由（MoE）

MoE 基本原理与优势



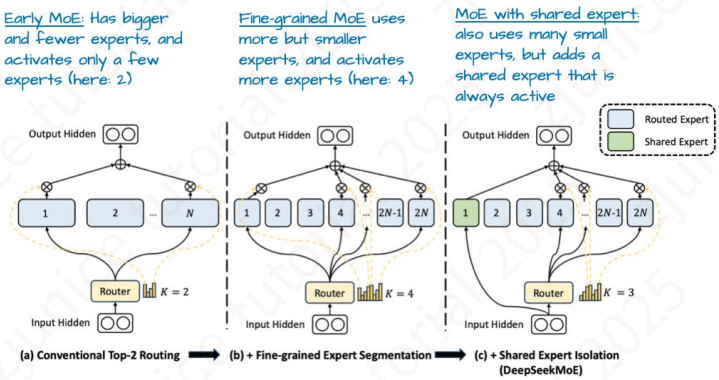
- **核心思想**：每层包含多位 FFN 专家，仅**激活少数专家**处理当前 token，计算与显存按激活比例缩放。
- **路由方式**：门控 $g(x)$（Top-k/Switch/Hash 等）根据表示选择专家；常见损失含**负载均衡**与**熵约束**，避免“热门专家”拥塞。
- **关键超参**：专家数 E、激活数 k、容量系数（Capacity Factor）、是否**共享专家**；这些决定**能力—吞吐—稳定性**三者平衡。

DeepSeek MoE 架构设计



- **专家编制**：每层 256 路由专家 + **1 个共享专家**；推理时共享专家常驻，提升**基础能力与鲁棒性**。
- **以小换多（Small-Experts）**：将单专家 MLP/FFN 宽度 $d_{\text{ff}}^{(e)}$ 明显小于稠密基线 d_{ff}，从而降低**单专家显存与算子尺寸**；
- **多专家高覆盖（$E\uparrow$）**：每层配置 E=256（+ 共享），以“**更多数量 × 更小体积**”替代“少而大”，提升**语义覆盖与专家专化度**

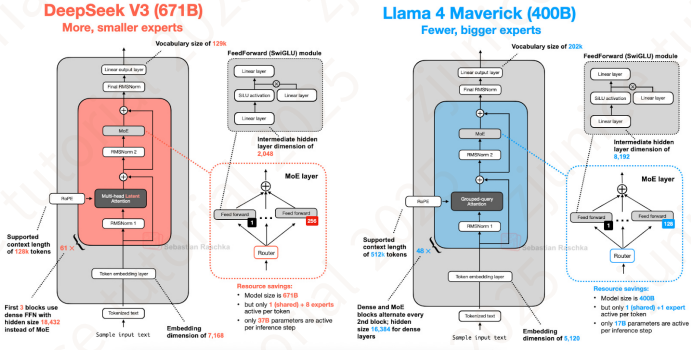
DeepSeek：训练与推理策略



- **冷启动**：先用较大 τ（高温）与较小 k 训练，使负载更均匀，再逐步**降温/增** k 提升分工。
- **容量计划**：warmup 阶段 cap_factor↑防丢弃；稳定后逐步调回目标值，平衡**吞吐与质量**。
- **推理调度**：批内重排（把路由到同专家的 token **聚簇**）+ **节点优先**分配；与分页/块化 KV 协同。

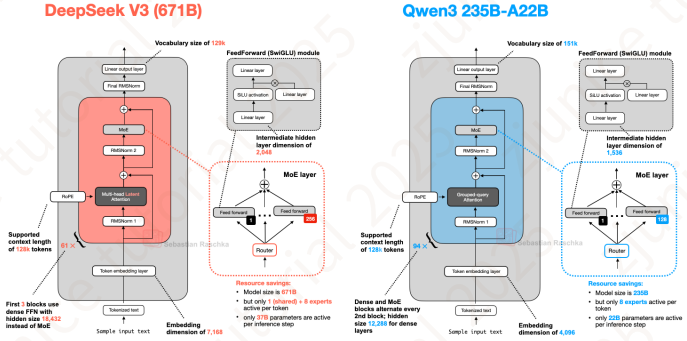
Llama 4 与 DeepSeek V3：MoE 架构对比

- **层级布局**：DeepSeek 近似“层层 MoE”；Llama 4 采用**交替 Dense/MoE**。
- **激活数**：DeepSeek k=8（+ 共享）；Llama 4 k=2（+ 共享），**通信与调度**更轻，延迟更稳。
- **结果**：相同吞吐预算下，交替布局能**更稳定上线**。



Qwen3 MoE 架构创新

- **无共享专家**：强调专家**完全分工**与**专业化**，减少共享路径对路由学习的“短路”影响。
- **专家与激活**：每层设置固定数量专家?。
- **均衡目标**：使用**全局批次负载均衡**损失，结合**容量回退**避免 token 丢弃，提高**有效样本利用率**。



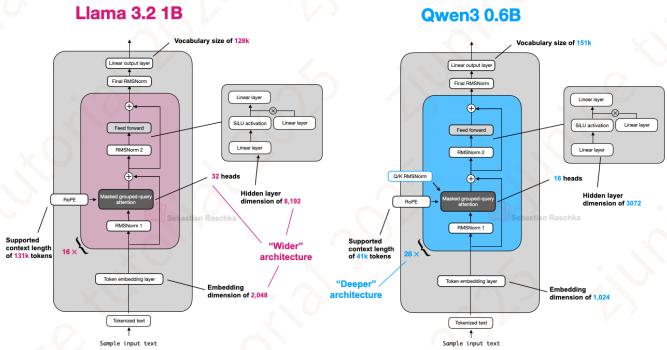
总结：稀疏激活与专家路由（MoE）

模型	专家组织	激活专家	共享专家	路由/均衡	上线时间
DeepSeek V3	256+1 共享	8+1	有	负载均衡优化	2024-12-27
LLaMA 4	16+1 共享	2+1	有	交替分布	2025-04-05
Qwen3	128	8	无	低成本路由	2025-04-29

参数组织与任务分流

Qwen3 Dense 架构创新

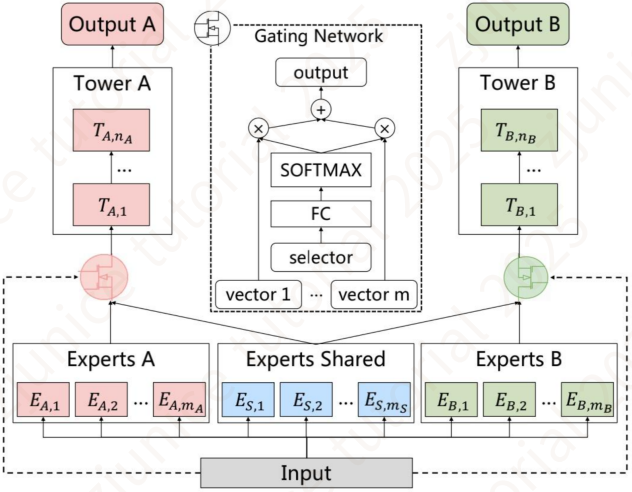
- **FLOPs 粗算**：单层注意力/FFN 主项 $\propto d^2$；在预算固定时，**加深（增层数 L）比加宽（增 d）更划算**。
- **稳定训练**：残差缩放 $1/\sqrt{L}$、分层学习率、增大归一化权重，避免深层梯度退化。
- **迁移**：深层结构对蒸馏/量化更稳，端侧显存更可控。



	Tokens/sec	Memory
Llama3Model 1B	42	2.91 GB

	Tokens/sec	Memory
Qwen3Model 0.6B	24	1.49 GB

Gemma 3n：PLE 让 LLM 落地边缘设备

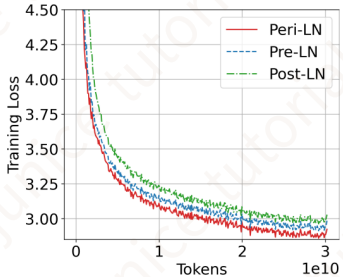


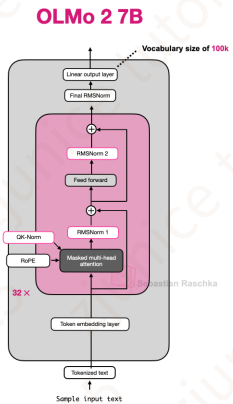
- **结构**：Shared Layers → **Coarse Branch** → Task Experts（**无 All-to-All**），**梯度全额回传**。
- **为何胜过 MoE（端侧）**：无跨分片通信，**延迟可预测**；参数可**按分支裁剪**，显存/功耗友好。
- **任务分流**：粗分支先聚类语义，后期专家**按任务域细化**（如 ASR/NLU/翻译等），避免**过早专业化**。
- **落地经验**：端侧建议**少量分支 + 小专家**，配合**分支感知蒸馏/量化**；云端可在此基础上**蒸馏到 Dense/MoE**。

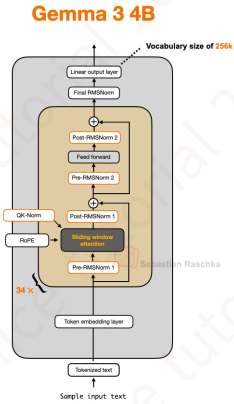
归一化与优化器

Gemma 3 的混合归一化机制

- Pre-Norm 稳定但**信号幅度易漂移**；Post-Norm 可**抑制激活爆炸**但深层可能梯度变弱。
- 在 Attention/FFN 的**前后均加 RMSNorm**，**前稳梯度、后稳幅度**，降低深层训练抖动。
- **效果**：长深度训练更稳，对指令跟随与多任务迁移**更鲁棒**。







AdamW 与 Muon 优化器原理与区别

AdamW 优化器更新公式:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$
$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t}$$
$$\theta_{t+1}^{(i)} = \theta_t^{(i)} - \eta \cdot \frac{m_t^{(i)}}{\sqrt{v_t^{(i)}} + \epsilon}$$

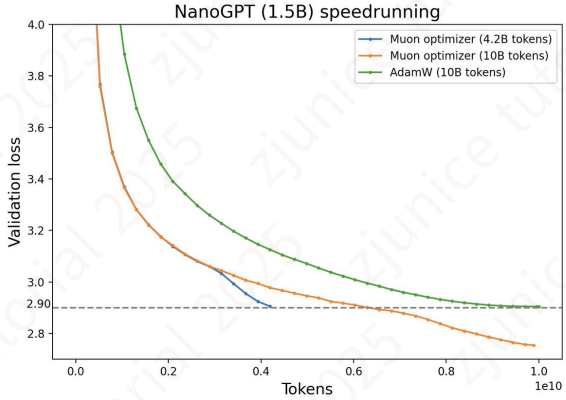
Muon 优化器更新公式:

$$\|m_t\| = \sqrt{\sum_{i=1}^{n} (m_t^{(i)})^2}$$
$$\theta_t = \theta_{t-1} - \eta \cdot \frac{m_t}{\|m_t\| + \epsilon}$$

其中 g_t 为梯度，θ_t 为当前参数，η 为学习率。

- **AdamW**：逐参数自适应优化，稳定但尺度感知差，容易震荡
- **Muon**：在梯度更新时对权重使用 L2 范数进行归一化，使参数更新方向一致，防止不同层尺度失衡
- 梯度更新从“标量维度”提升为“向量级别”归一
- 更适用于深层/宽层大模型，收敛快、更鲁棒
- Kimi 2 提升稳定性的核心关键之一

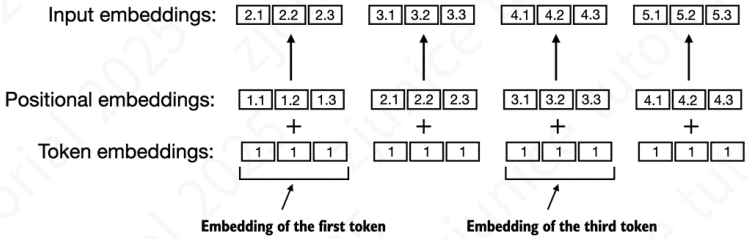
Kimi 2：Muon 优化器应用表现与优势



- 同样训练 Token 下，Muon 优化器 Loss 下降更快
- 更少的训练步骤获得更低的验证 Loss，节省算力和训练时长
- 更快收敛 = 更短训练时间，更适合大规模 LLM 快速迭代

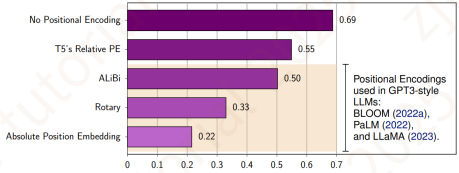
位置编码的演进

SmolLM NoPE：背景与基本原理



- **理念**：完全移除显式位置编码；仅用**因果 Mask** 保序，模型从数据中**隐式学习顺序线索**。
- **优势**：无“外推失真”（不依赖 RoPE 频率外推），**长序列泛化**更稳；参数/实现更简单。
- **边界**：对**强顺序/算子化**任务（代码/数学）可能吃亏；可在**高层混入少量相对编码**作折中。

SmolLM NoPE 的优势与实验效果



Context Length	Model Variants	begin	needle	context	qc
8k	RoPE	0.3863	0.0328	0.3809	0.2000
	QK-Norm	0.0242	0.0173	0.8020	0.1565
	NoPE	0.3058	0.0454	0.4501	0.1987
32k	RoPE	0.3541	0.0201	0.4343	0.1915
	QK-Norm	0.0064	0.0056	0.8517	0.1364
	NoPE	0.2807	0.0325	0.4981	0.1886
128k	RoPE	0.3463	0.0010	0.4751	0.1776
	QK-Norm	0.0010	0.0004	0.8993	0.0994
	NoPE	0.0846	0.0073	0.8156	0.0925

- text 泛化能力突出：NoPE 对长序列任务具有极强泛化优势
- text 小模型表现优异：对小/中等规模模型，能提升极长文本检索任务表现
- text 任务依赖性：主流任务如 MMLU、CommonsenseQA 等，NoPE 与 RoPE 表现接近
- text 训练简单，参数少：结构极简，无需位置参数，减少显式超参调优

Model	Val Loss	MMLU	HellaSwag	CommonsenseQA	ARC-E	ARC-C	Needles 65k
RoPE	1.52	48.55	73.74	68.30	81.05	39.13	9.82
QK-Norm	1.53	48.21	73.68	68.23	80.54	38.98	7.93
NoPE	1.58	47.61	72.16	66.42	76.94	37.12	9.03

架构总结与趋势展望

架构总结与趋势展望

- **MoE 架构仍是核心突破口**
  - 参数利用率低 → 稀疏激活 + 共享专家成为共识
  - DeepSeek 和 LLaMA4 展现出不同的专家组织策略，分别追求更强能力 vs 更优吞吐比
  - 路由机制仍存在不稳定与负载不均问题，是未来研究重点
- **KV Cache 优化成为 Attention 核心议题**
  - MLA 与 GQA 分别强调压缩效率与工程兼容
  - 推理与上下文长度提升驱动结构创新
  - MLA 提升压缩效率但部署复杂，GQA 兼容性更强
- **归一化与优化器设计有广阔空间**
  - RMS 混合归一化增强训练稳定性
  - Muon 优化器展现更强收敛与尺度适应
- **端侧优化推动结构精简与任务分流**
  - NoPE、PLE 等技术降低部署门槛
  - “轻量模型 + 高性能”逐渐成为焦点

Thank you

Networked Intelligence for Comprehensive Efficiency (NICE) Lab
College of Information Science and Electronic Engineering
Zhejiang University
https://nice.rongpeng.info