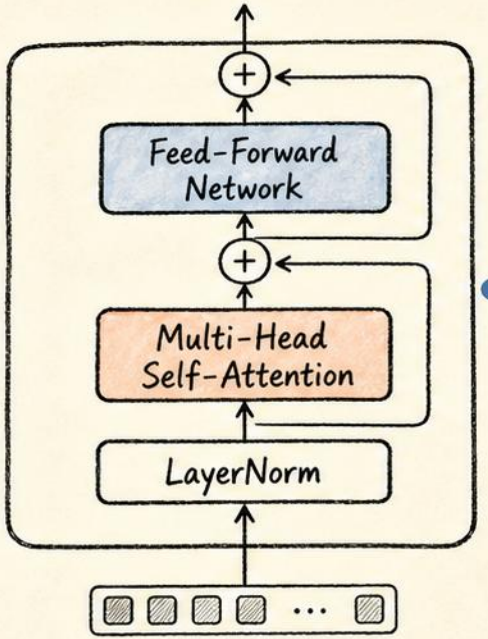
Speed Always Wins 解读

高效 LLM 架构地图

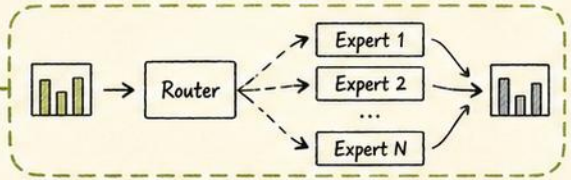


Transformer

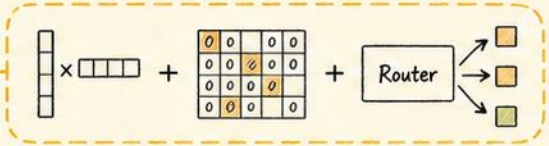
Linear

Sparse

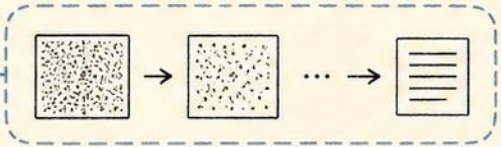
MoE



Hybrid

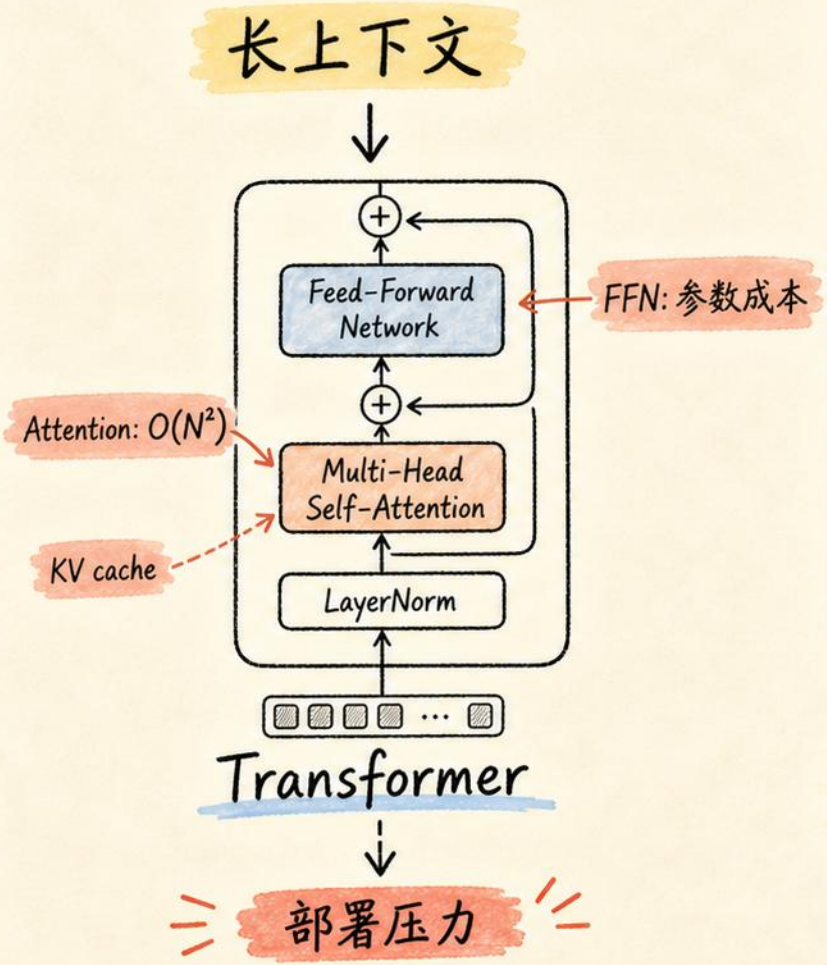


Diffusion



arXiv 2508.09834

效率变成硬约束



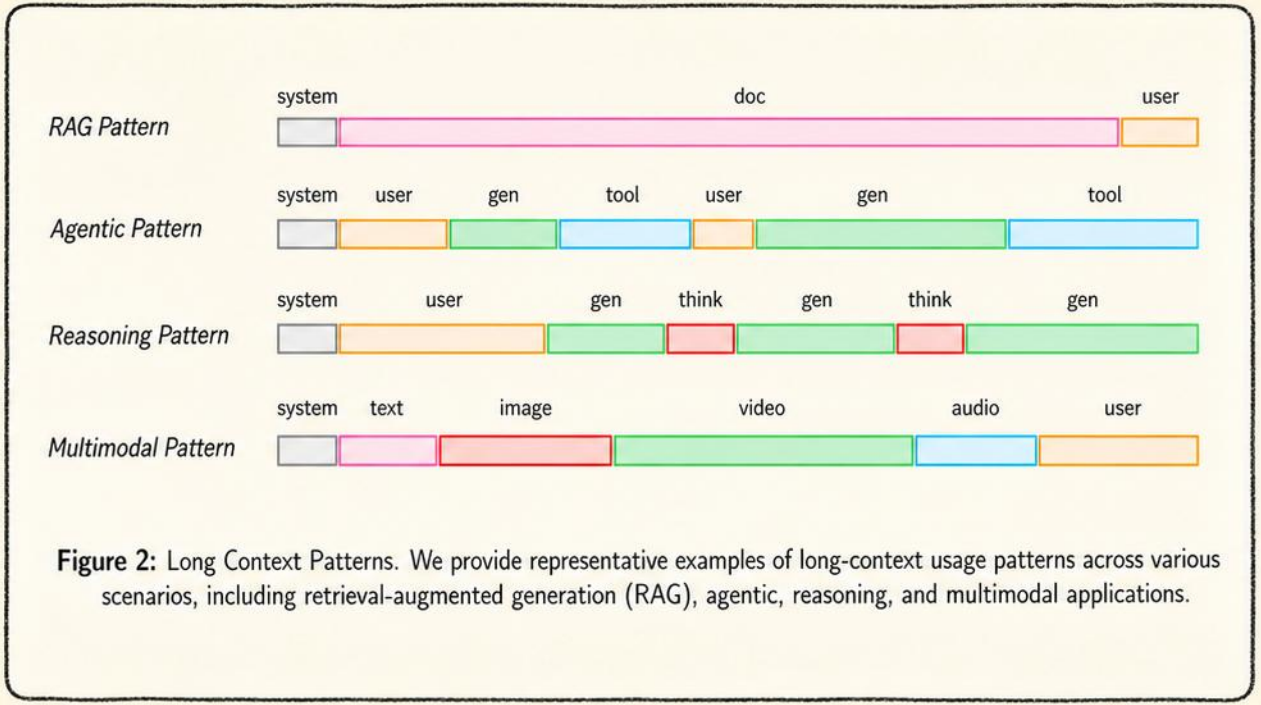


Figure 2: Long Context Patterns. We provide representative examples of long-context usage patterns across various scenarios, including retrieval-augmented generation (RAG), agentic, reasoning, and multimodal applications.

RAG | Agent | Reasoning | Multimodal

☆ 七条效率路线

① Linear

② Sparse

③ Full Attention

④ MoE

⑤ Hybrid

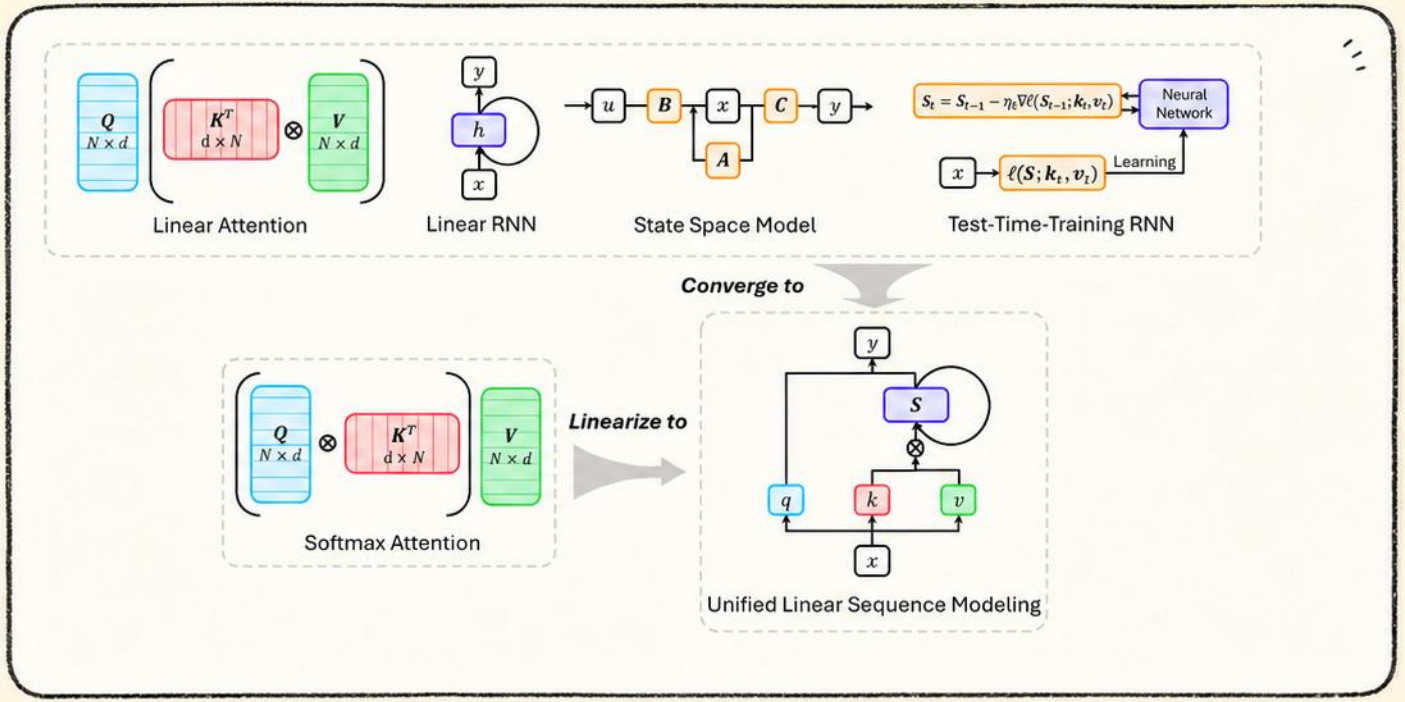
⑥ Diffusion

⑦ Multimodal

Figure 3: A Comprehensive Taxonomy of Efficient Architectures for Large Language Models.

☆ 不是单点技巧，是架构地图 ☆

把注意力写成状态



状态读写

- ✓ 把全局信息压缩到状态 S
- ✓ 读写以 q, k, v 与状态交互

递推更新

- ✓ 一步步更新状态 S
- ✓ 从全局计算走向递推计算

查询读出

- ✓ 用查询 q 从状态读出 y
- ✓ 高效输出所需信息

$O(N^2) \rightarrow O(N)$

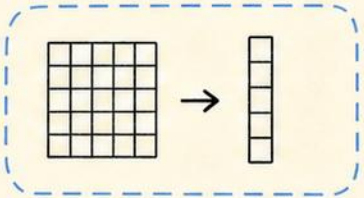
少 KV cache

→ Softmax → Linear / Recurrent

从零训练高效架构太贵，
linearization 试图把已有 Transformer 低成本转换成 linear/recurrent/hybrid 模型。

低成本适配

替换注意力



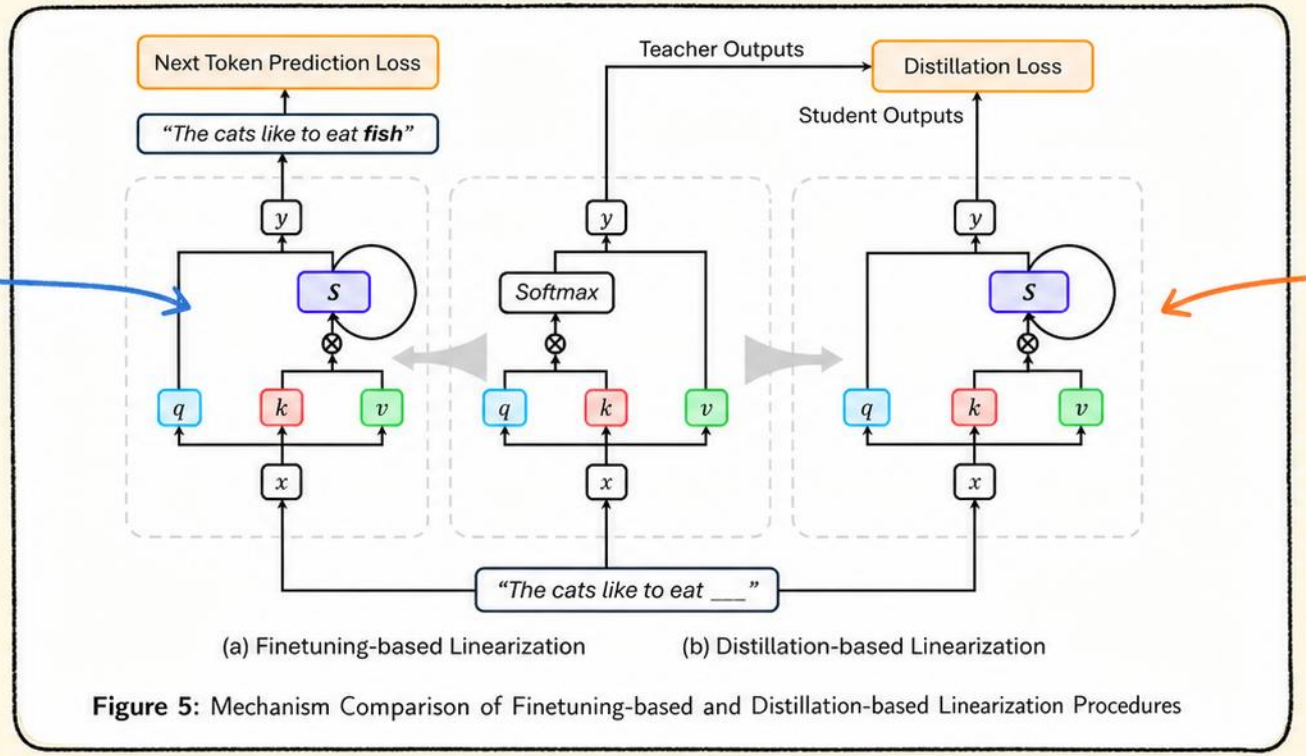
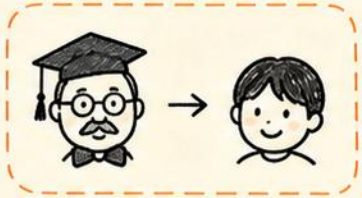
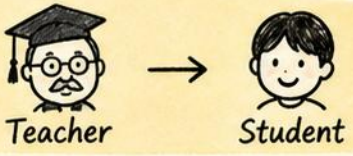


Figure 5: Mechanism Comparison of Finetuning-based and Distillation-based Linearization Procedures

蒸馏迁移



swap → finetune



teacher → student

不是所有 token 都互看

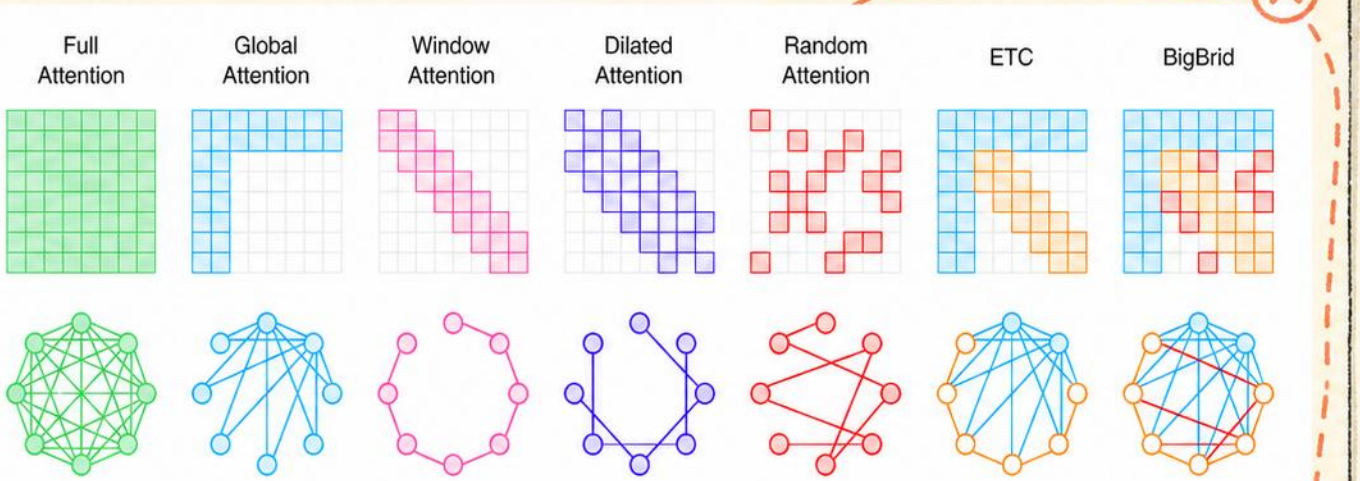
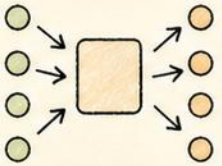
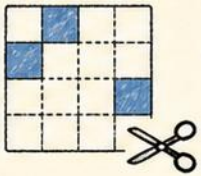


Figure 7: Example Patterns of Static Sparse Attention. ETC and BigBrid are representative examples of mixed static sparse attention mechanisms, which combine multiple fixed sparsity patterns within a single attention framework.

固定稀疏

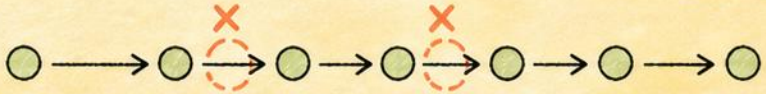


动态路由



推理剪枝

少算关键边

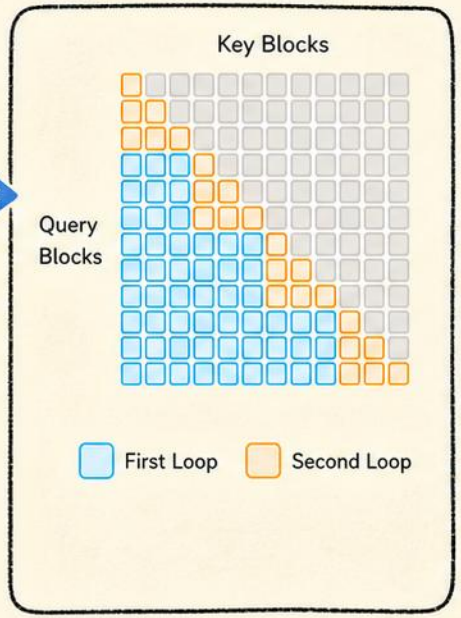


精确注意力也能更快

保留 softmax

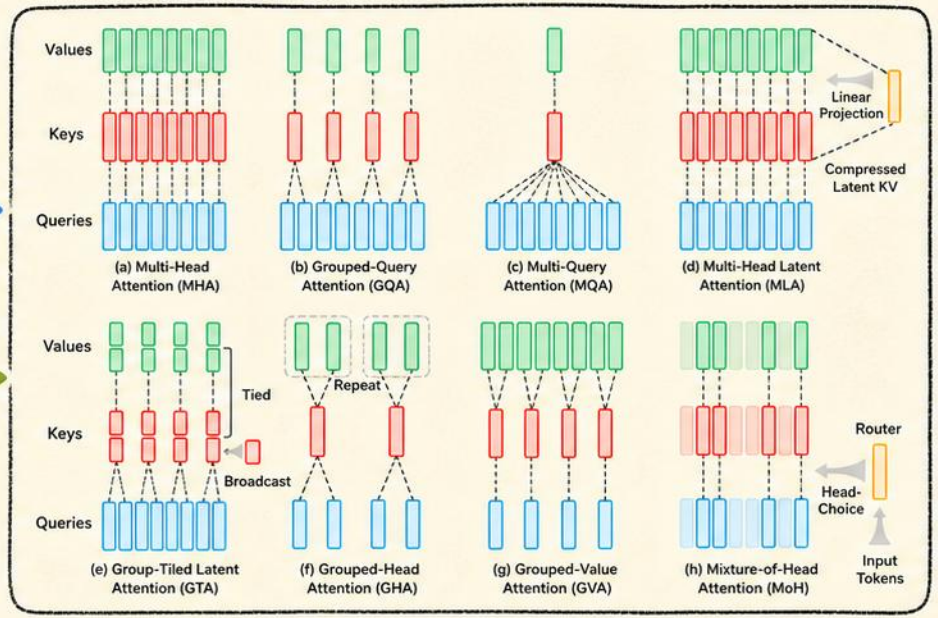
IO-aware

高 IO 开销

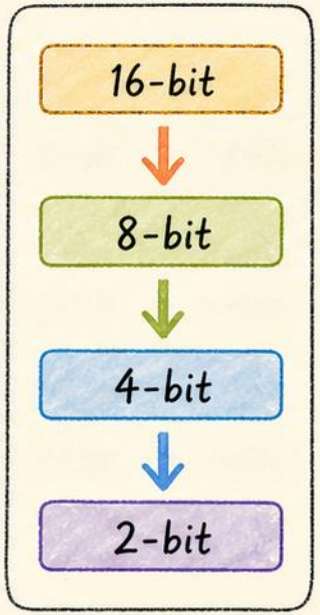


KV 压缩

KV Cache



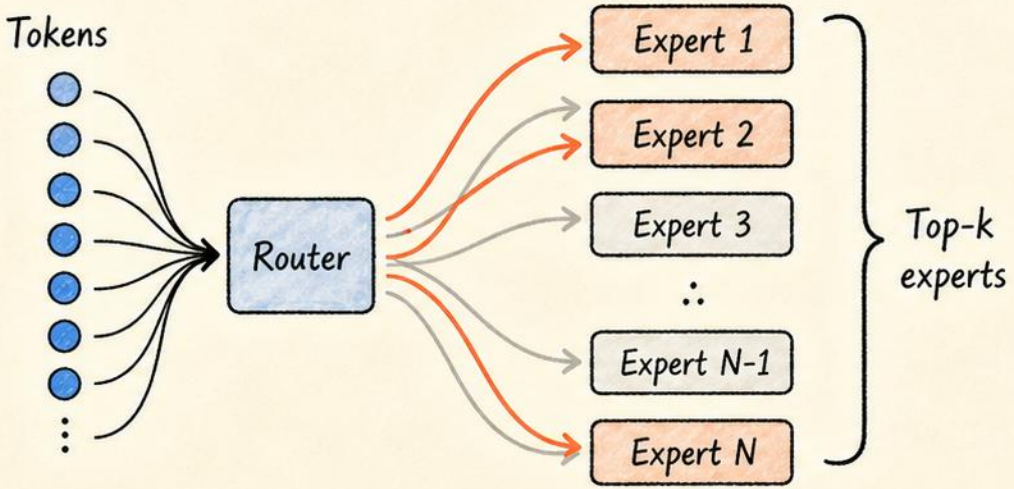
低精度



从 IO-aware tiling、head/KV sharing、低精度量化入手，让“精确注意力”更快。

MoE 的核心是 conditional computation

容量大，激活少



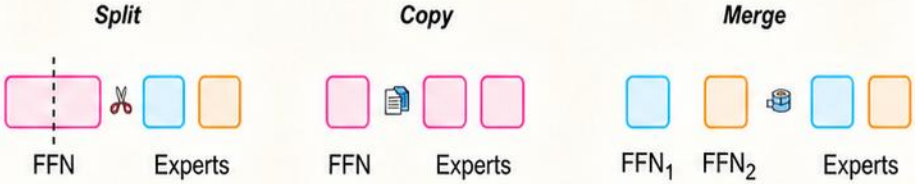


Figure 12: MoE Conversion Strategies.

Split：将原始 FFN 拆分，得到多个专家，减少单个专家规模。

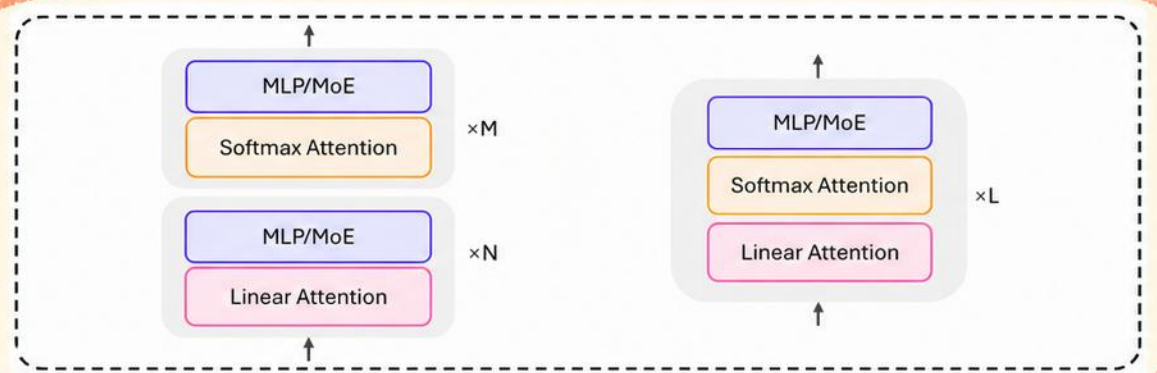
Copy：复制 FFN，直接得到多个同构专家，简单高效。

Merge：合并多个 FFN，形成多个更强的专家，提升表达能力。

capacity != compute

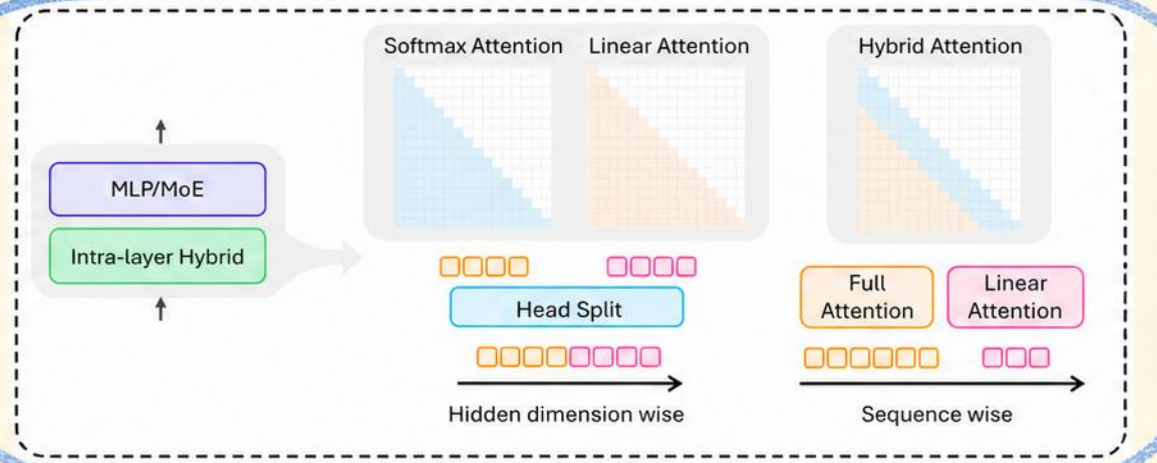
⚠ 负载均衡 | ⚠ 通信开销

层间混合



(a) Inter-layer Hybrid

层内混合



(b) Intra-layer Hybrid

Linear
效率

Softmax
精确召回

效率层 + 精确层

跳出逐 token 解码

Diffusion LLM 把文本生成看作 masked/denoising 过程，换来并行解码和更强可控性，但仍需解决固定长度,质量和 AR 兼容问题。

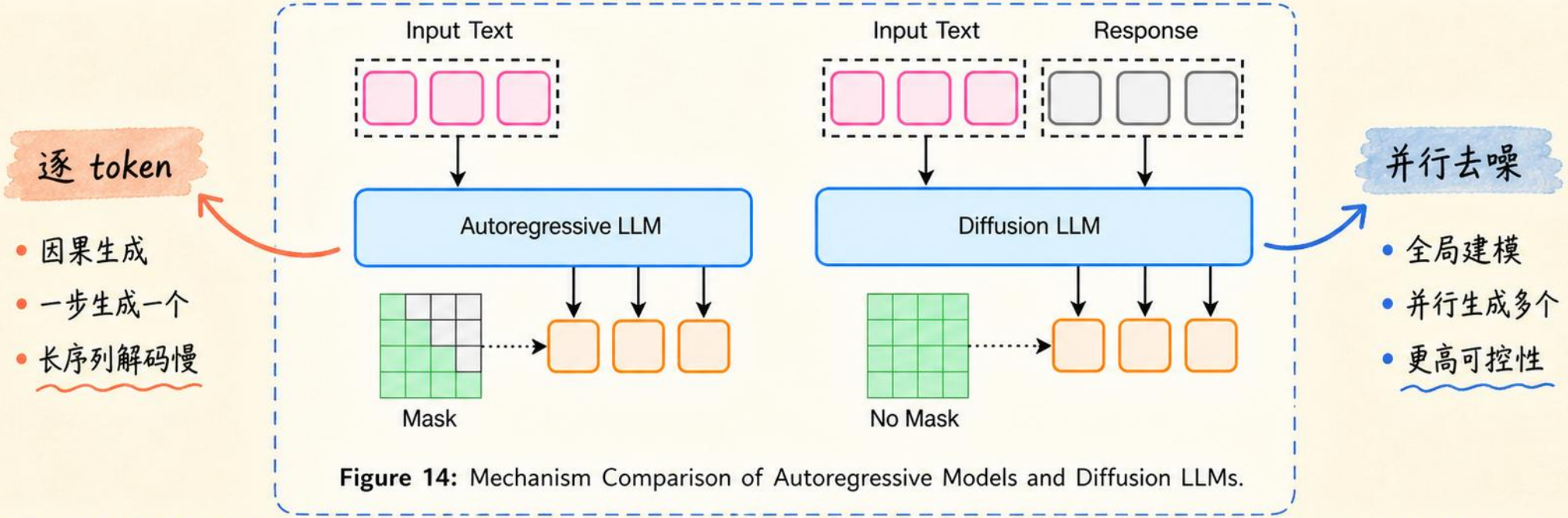


Figure 14: Mechanism Comparison of Autoregressive Models and Diffusion LLMs.

可控生成

更强的全局条件与风格控制

固定长度?

输出长度如何自适应?

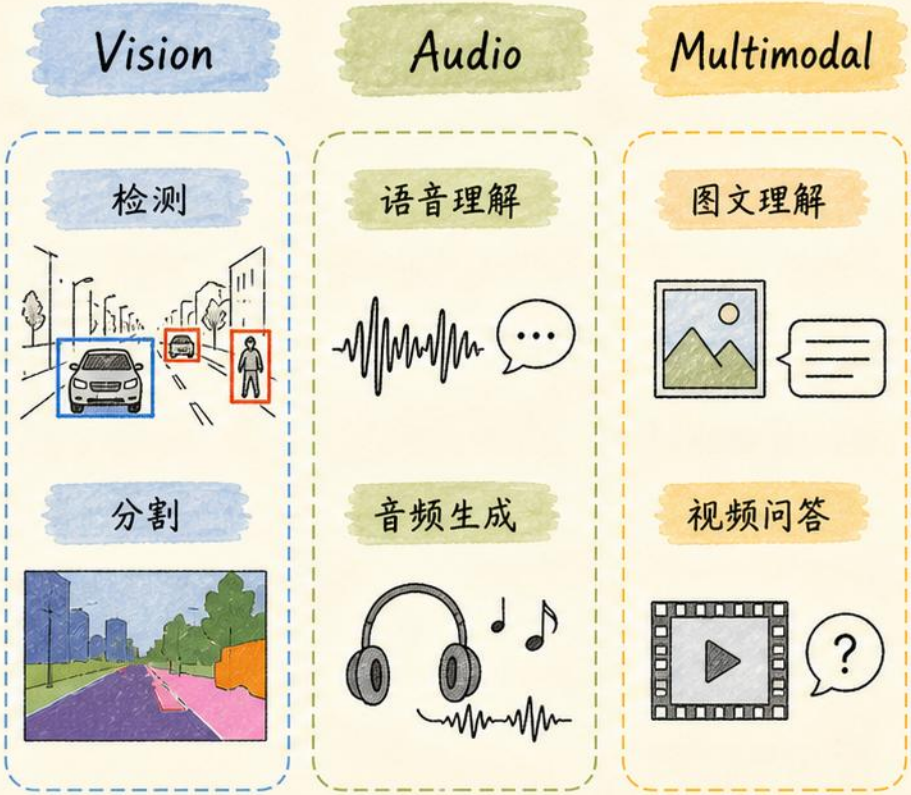
质量?

如何进一步提升生成质量?

同一套原则外溢

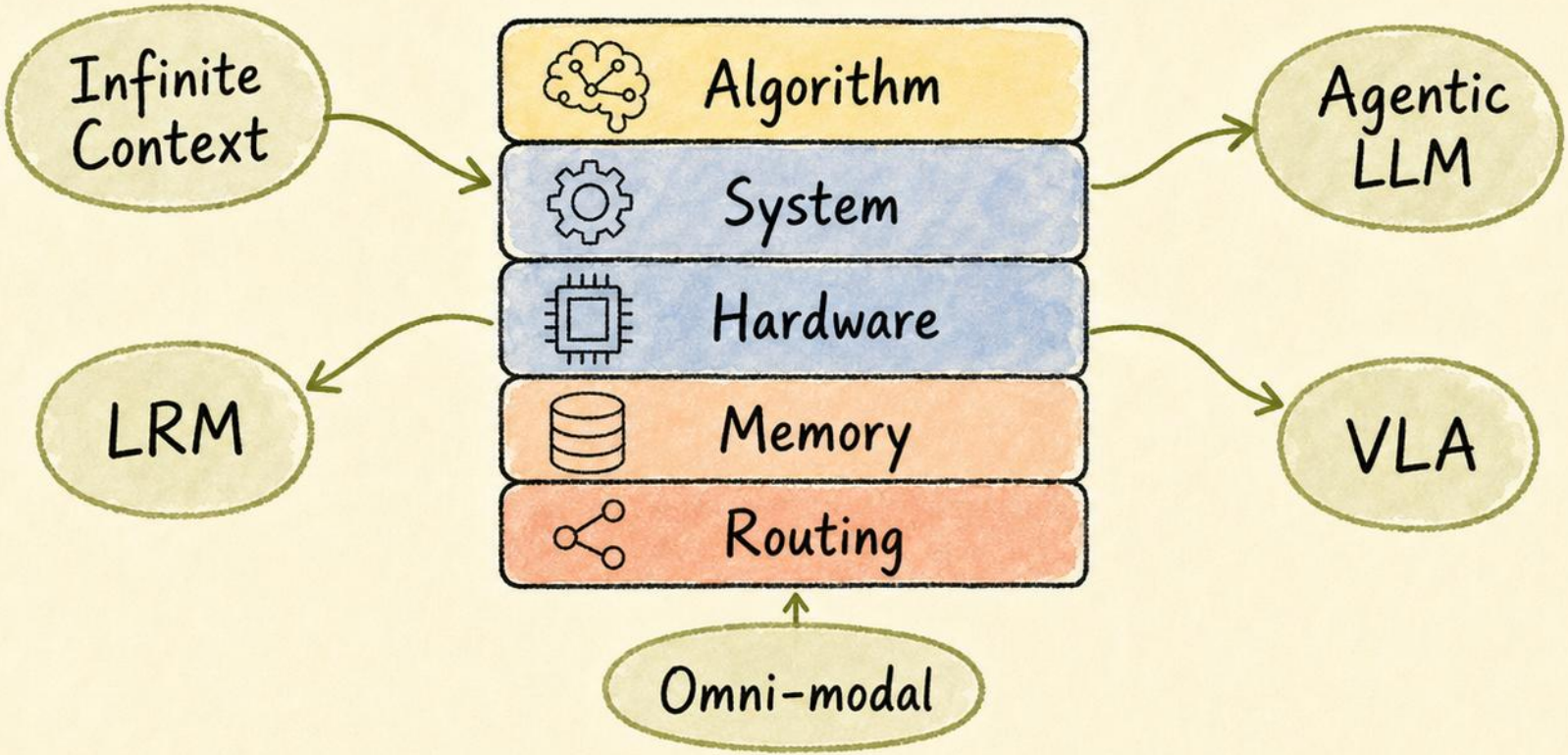
Table 2: Overview of Applications of Efficient Architectures Across Modalities and Categories.

Modality	Category	Approach
Vision	Classification	InsectMamba [208], V-MoE [332], MoE-CNN [368], pMoE [369], Res-vmamba [370], Mammil [371], Spectralmamba [372], Memorymamba [373]
	Detection	Vig [205], Vision-rwkv [206], Tutel [207], Mamba yolo [374], Voxel mamba [209], Mim-istd [375], Htd-mamba [376], Soar [377]
	Segmentation	RWKV-SAM [378], Segman [379], VM-Unet [380], Pyramidmamba [381], Vig [205], Vision-rwkv [206]
	Enhancement & Restoration	Restore-rwkv [382], Dvmsr [383], Q-mambar [384], RWKV-IR [385], Pixmamba [386], Water-mamba [387], Retinexmamba [388], Llemamba [389], Fouriermamba [390], Vmambair [391], Serpent [392], Matir [393], Cu-mamba [394], Lfmamba [395]
	Generation	DiS [396], Dim [397], DiM [210], Zigma [398], AiM [399], Maskmamba [400], Dimba [401], DiM-3D [402], Gamba [403], Diffusion-rwkv [404], Sdit [405]
	Medicine	U-mamba [211], Vm-unet [212], Rwkv-unet [213], Segmamba [406], Zig-rir [407], Bsbp-rwkv [408], Mambamil [409], Mmr-mamba [410], Mambamir [411], Vmambamorph [412], I2i-mamba [413], Delta-wkv [414]
	Autonomous Driving	Mambabev [214], Occrwkv [415], H-mba [416], Trajectory mamba [417], Drama [418], Salm^2 [419]
	Remote Sensing	RS-Mamba [420], Samba [421], Changemamba [422], Rs3mamba [423], Hsimamba [424], Pan-mamba [425], LE-Mamba [426], Rsdehamba [427], FMSR [428], Rscama [429]
Audio	Understanding	Audio mamba [215, 430, 431], Mamca [216], Rawbmamba [217], BiMamba [220], Ssamba [432], Dual-path mamba [221], Spmamba [222], VAD [223]
	Enhancement & Generation	SEMamba [433], Tramba [434], SaShiMi [218], Music-Diff [219], oSpatialNet-Mamba [435]
Multimodality	Understanding	MaTAV [224], Avs-mamba [225], Av-mamba [436], VisualRWKV-UHD [226], Rwkv-clip [437], Lavida [203], LiMoE [438], Uni-MoE [439], VL-MoE [227], Moe-llava [228], MoCLE [229], Llava-mole [230], PaCE [440], Llada-v [201], Dimple [366]
	Unified	Fragkiadaki [202], Mmada [204]



长序列 + 稀疏计算 = 从文本到多模态

快是系统命题



架构、系统、硬件一起设计