目录

K2 架构设计与工程权衡

探索 Kimi K2 的 MoE 架构设计理念，了解在规模、效率和性能之间的精妙权衡

核心架构：1T 参数 MoE

Kimi K2 延续并优化了 MoE 架构，其核心思想是通过激活一小部分"专家"网络来处理信息， 从而在巨大模型规模下保持高效推理。

1.04T	32B	384	8 + 1
总参数量	激活参数量	专家数量	激活专家

架构特点详解

总参数量与激活参数

总参数量：1.04 万亿 (1.04T)
激活参数量：每个 token 推理时仅激活 32B 参数，实现了极高的计算效率

专家系统 (MoE)

- **专家数量**：384 个，体现了"更多、更小专家"的设计哲学，以实现知识的精细化分工
- **激活机制**：每个 token 会路由到 8 个选定专家和 1 个共享专家进行处理
- **MoE 引入**：仅首个 block 使用标准 FFN，从第二个 block 开始便引入 MoE 结构

MoE 架构动态演示

展示每个 Token 如何路由到 8 个选定专家和 1 个共享专家进行处理

开始演示 重置

模型 Block 结构

Block 1	Block 2	Block 3	Block 4	Block 5
标准 FFN	MoE 结构	MoE 结构	MoE 结构	MoE 结构

说明：仅首个 block 使用标准 FFN，从第二个 block 开始引入 MoE 结构

Token 输入与路由

Token → Router

专家系统 (384 个专家)

显示 24 个代表全部 384 个专家

E1	E2	E3	E4	E5	E6
E7	E8	E9	E10	E11	E12
E13	E14	E15	E16	E17	E18

E19 E20 E21 E22 E23 E24

共享专家： Shared Expert 待激活

384	0	0
总专家数	选定专家	共享专家

词汇表与上下文

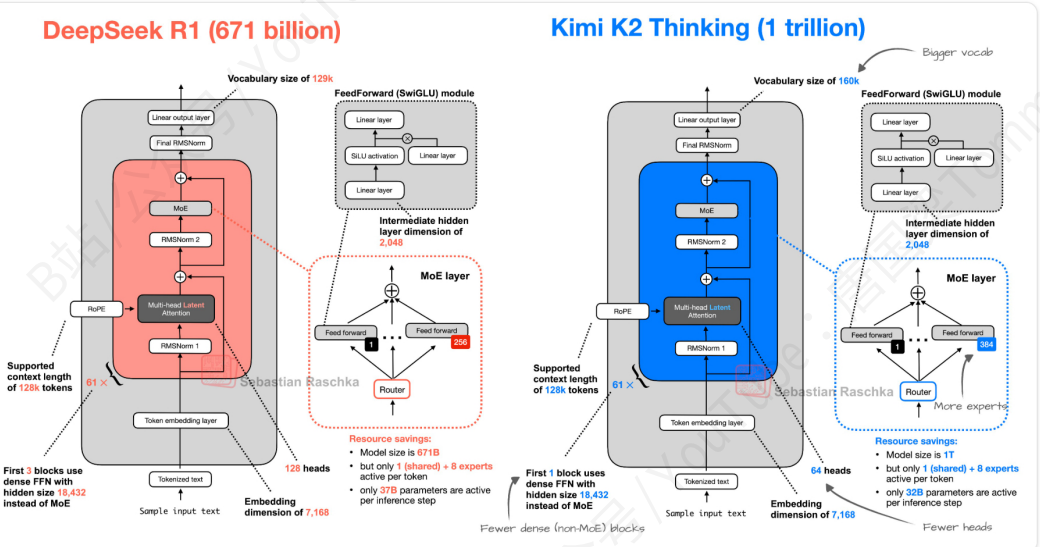
词汇表：160k，更大的词汇表能更高效地编码多语言和专业术语
上下文窗口：支持高达 256k tokens，能够处理和记忆数百页的文档内容

注意力机制

采用 **MLA (Multi-head Latent Attention)** 机制，通过 latent 压缩空间减少 KV cache 占用，提升长上下文推理稳定性

KIMI K2 vs DeepSeek R1

通过对比 Kimi K2 和 DeepSeek R1，可以看出 Kimi K2 在多个关键维度上做出了独特的设计选择：



技术特征	Kimi K2 Thinking	DeepSeek R1	设计意图分析
总参数量	1万亿 (1T)	6710亿 (671B)	追求更高的模型容量和性能上限
词汇表大小	160k	129k	更高效地编码多语言文本和特殊符号，减少 token 数量
注意力头数	64 个	128 个	降低注意力计算复杂度和显存占用，加快推理速度
MoE 专家数量	384 个	256 个	"更多、更细颗粒度专家"实现知识稀疏化与专业化
激活参数量	32B	37B	推理时真正动用的算力更低，降低运行成本
非 MoE 层数	第 1 个 block	前 3 个 block	更早引入 MoE 结构，进一步优化资源利用

设计亮点

K2 的架构设计充分体现了"在性能和效率间精妙权衡"的工程哲学。通过减少注意力头、增加专家数量、更早引入 MoE 等创新， 实现了在万亿参数规模下的高效推理。

← 上一章：KIMI K2 简介

下一章：K2 预训练 →