2025了，如何回答“为什么现在的大模型都是decoder-only的架构”

去年，这个常见面试问题的回答都会参考大佬@Sam聊算法 在知乎的帖子。

今年，随着deepseek r1推理模型一战封神，强势颠覆openai在该领域的霸主地位，更多开源大模型也不甘示弱，更新速度好像做了火箭，于是，我重新整理了一下这个面试题的回答：

2024-2025 年几乎所有主流 LLM 都回归或延伸自 decoder-only：

- **表达能力**：Decoder-Only模型的自回归注意力矩阵为严格下三角形式并含单位对角线，**在理论上保持满秩**。Encoder-Decoder结构可能破坏注意力矩阵的满秩性，**潜在限制了模型性能上限。**
- **工程角度**: Decoder-only 的 KV-Cache 机制天然适配流水线并行和显存优化（如 vLLM 的 PagedAttention）。Megatron-LM、FlashAttention 等底层优化均优先支持因果（Causal）路径。MoE、量化、蒸馏等技术在decoder-only结构上更易实现。
- **预训练难度**：每一步都只看左侧信息，任务难度大，因此大模型+大数据下能逼出更通用的表征上限。
- **few-shot/zero-shot**：Prompt 在所有层都可注入梯度（隐式微调），比 Enc-Dec 两段式更直接。
- **隐式位置编码与外推优势**：Decoder-Only 将输入输出视为单一连续序列，仅依赖相对位置关系，无需显式对齐Enc-Dec的绝对位置索引。训练后可通过微调或插值轻松扩展上下文窗口（如 LongRoPE），而 Enc-Dec 需处理两套位置系统的兼容性问题。
- **多模态角度**: 主流方案（Gemini/GPT-4o）直接将视觉/音频 tokens 拼接至文本序列，由同一decoder处理，实现“早融合”的工程最优解。
- **轨迹依赖**：openai率先验证了该架构的训练方法和scaling law，后来者鉴于时间和计算成本，自然不愿意做太多结构上的大改动，就继续沿用decoder-only架构，迭代 MoE、长上下文、多模态。

像RWKV这种基于RNN的比较另类的模型结构，也有其适用场景，比如端侧小模型，但并非主流。