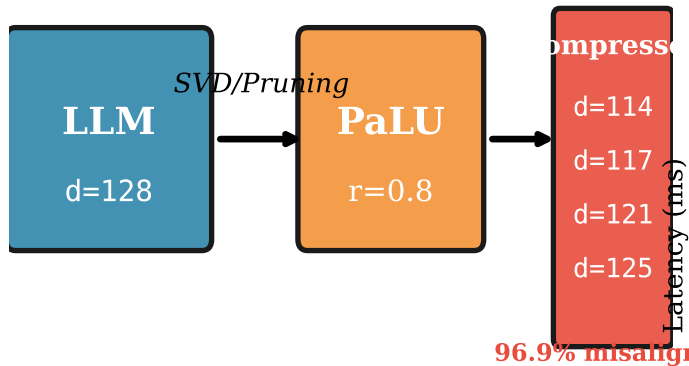


(a) Compression Produces Irregular Dimensions



GPU Alignment Requirements

FlashAttention: $d\%8=0$ | Tensor Core: $K\%16=0$ | float4: $K\%$

(b) SDPA Latency Cliff Effect

