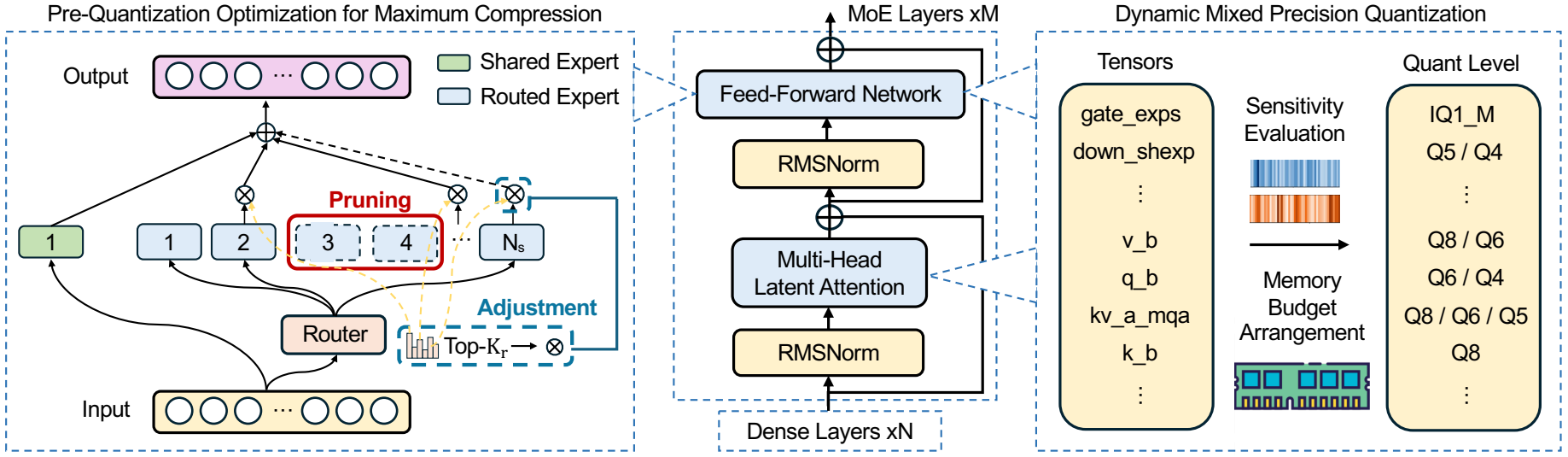
Pre-Quantization Optimization for Maximum Compression
Output
Shared Expert
Routed Expert
Pruning
1
1
2
3
4
$N_s$
Router
Top-$K_r$
Adjustment
Input
MoE Layers xM
Feed-Forward Network
RMSNorm
Multi-Head Latent Attention
RMSNorm
Dense Layers xN
Dynamic Mixed Precision Quantization
Tensors
gate_exps
down_shexp
v_b
q_b
kv_a_mqa
k_b
Sensitivity Evaluation
Memory Budget Arrangement
Quant Level
IQ1_M
Q5 / Q4
Q8 / Q6
Q6 / Q4
Q8 / Q6 / Q5
Q8