

Score-based eviction

H2O · SnapKV · R-KV · VaSE · TriAttention



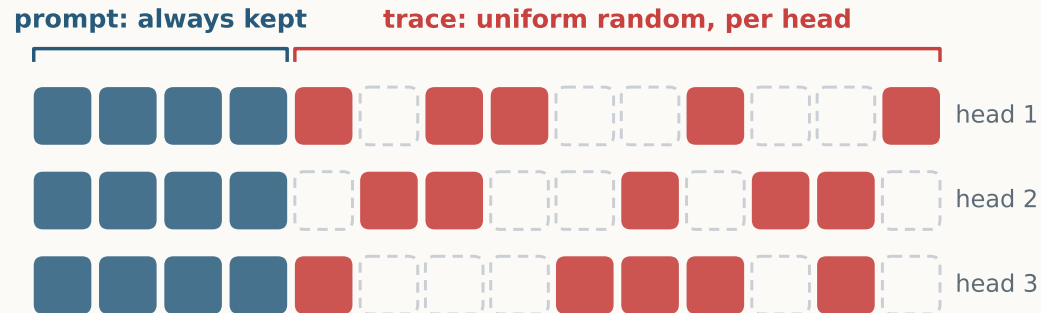
the question competes by score

TriAttention pins the whole prompt; the others keep only sinks

cost: a scoring pass at every eviction

Random Attention (ours)

no scores, no calibration, nothing to tune



every head keeps a different subset

so some copy of what the model still needs survives

cost: none — no scoring pass at all