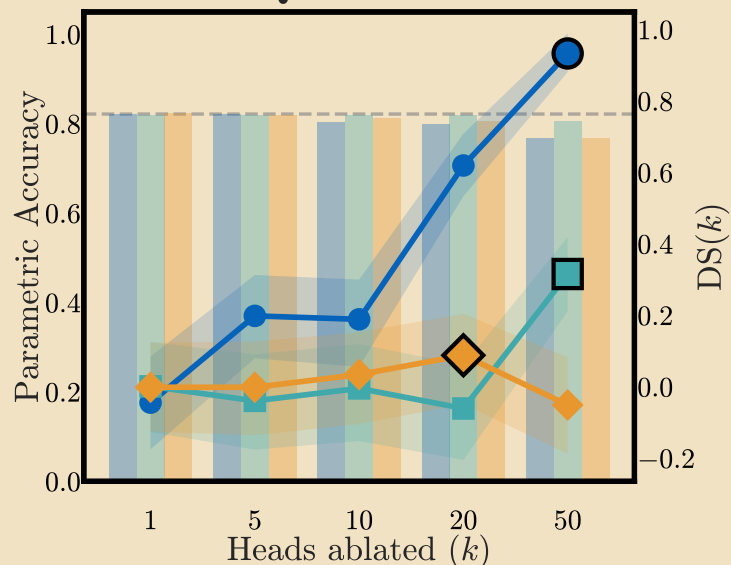
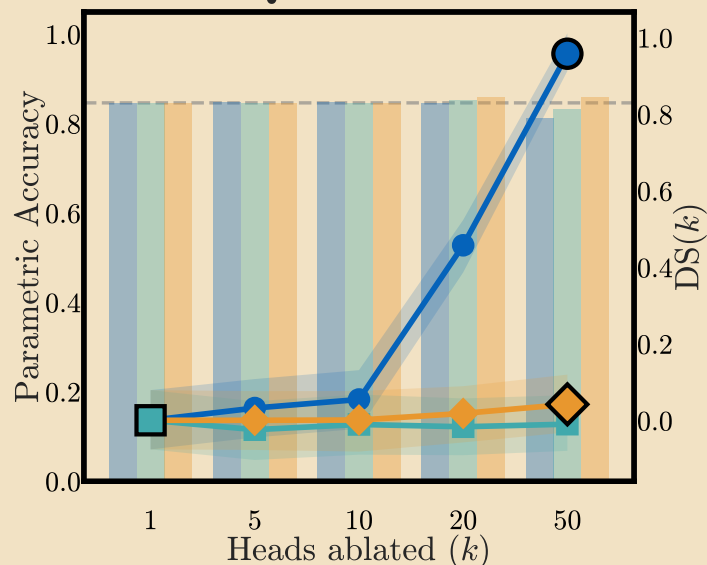


Dissociation Score: Logit-Contribution Top-k vs Bottom-k

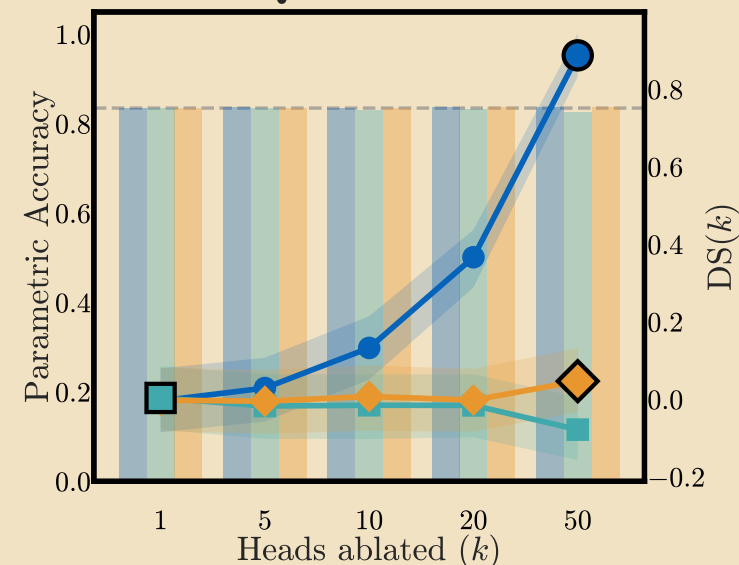
Qwen3-8B



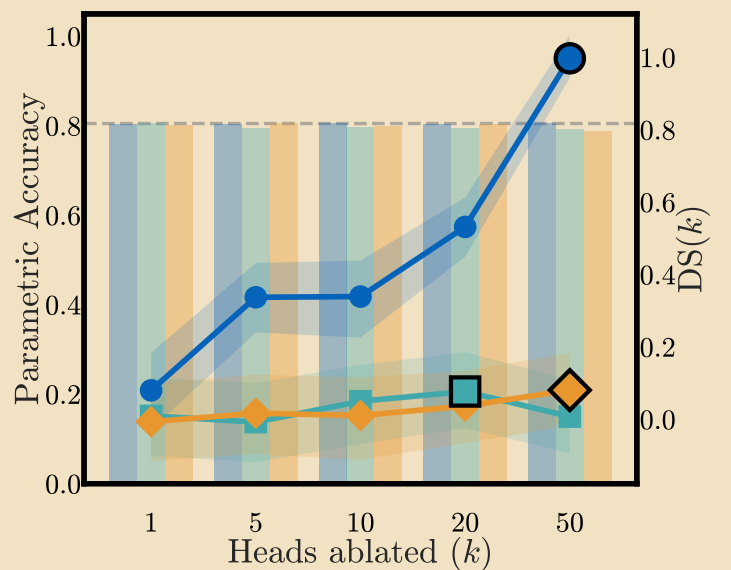
Qwen3-14B



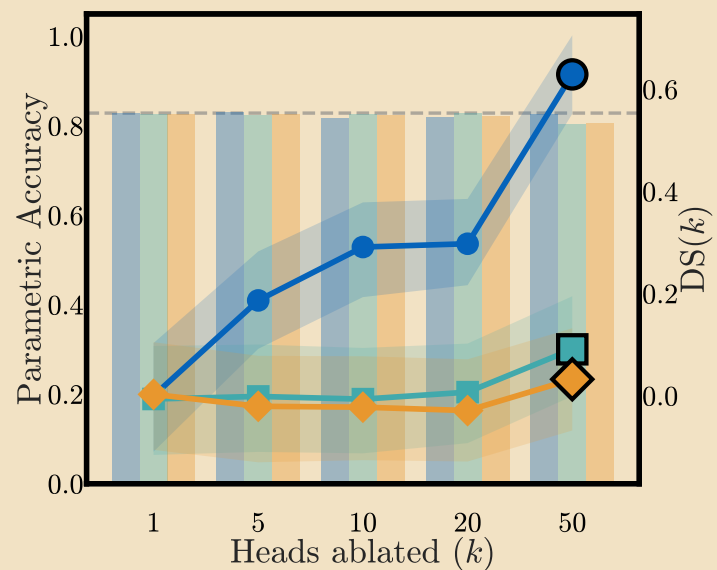
Qwen3-32B



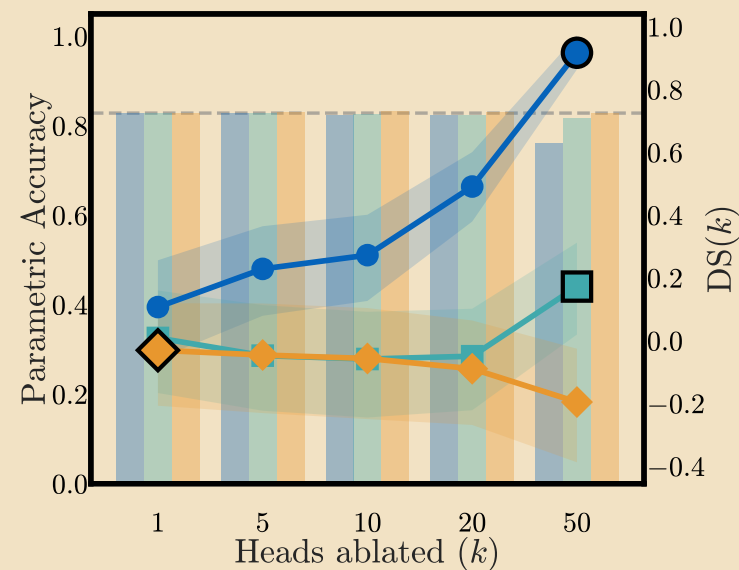
Gemma3-12B



Gemma3-27B



Olmo3.1-32B



--- Baseline (no ablation)

● LOCOS (Top-K)

● LOCOS (Bottom-K)

▲ Random