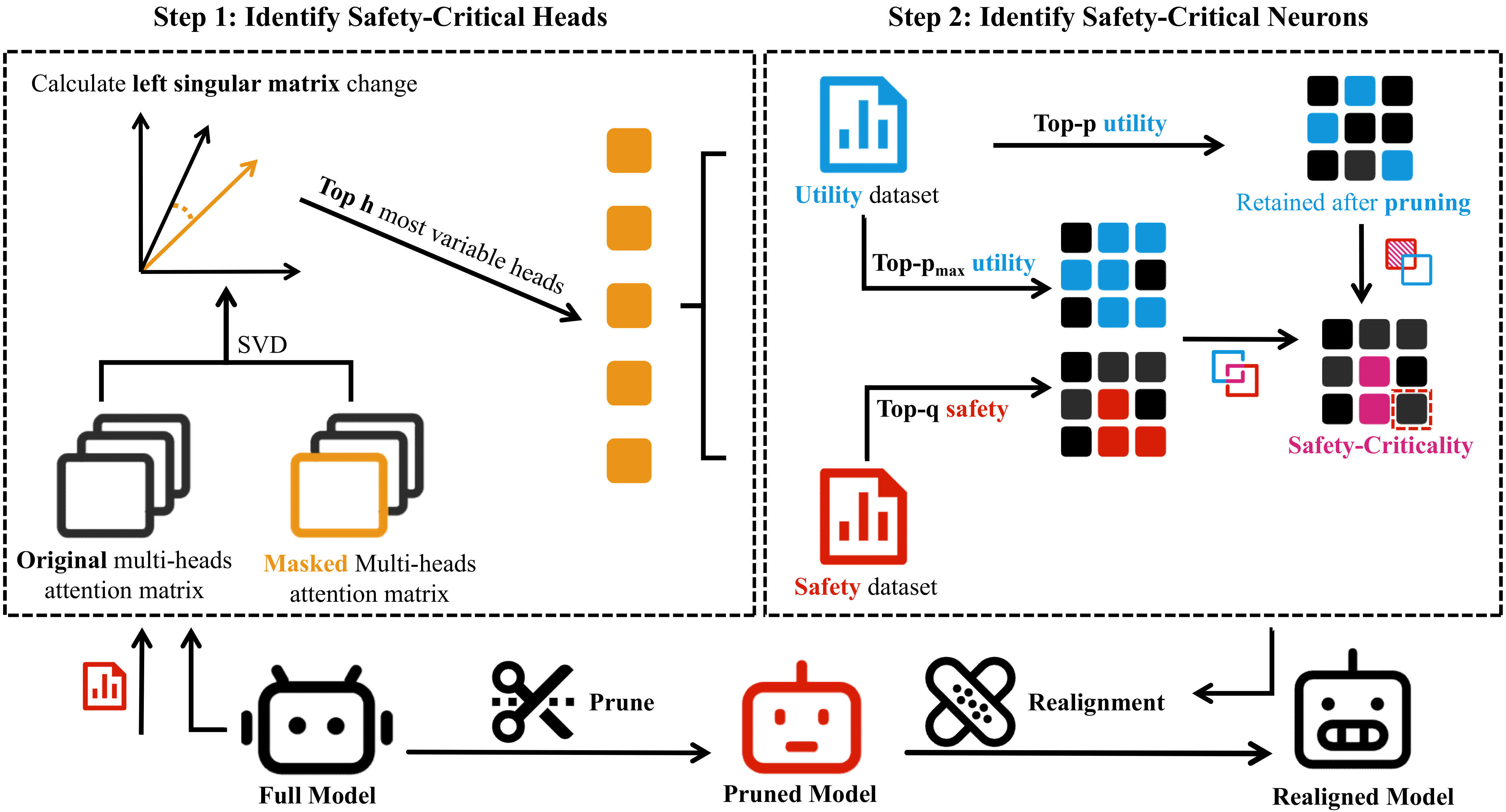
Step 1: Identify Safety-Critical Heads
Step 2: Identify Safety-Critical Neurons
Calculate **left singular matrix** change
Top h most variable heads
SVD
Original multi-heads attention matrix
Masked Multi-heads attention matrix
Utility dataset
Top-p utility
Retained after pruning
Top-p_{max} utility
Top-q safety
Safety-Criticality
Safety dataset
Prune
Realignment
Full Model
Pruned Model
Realigned Model