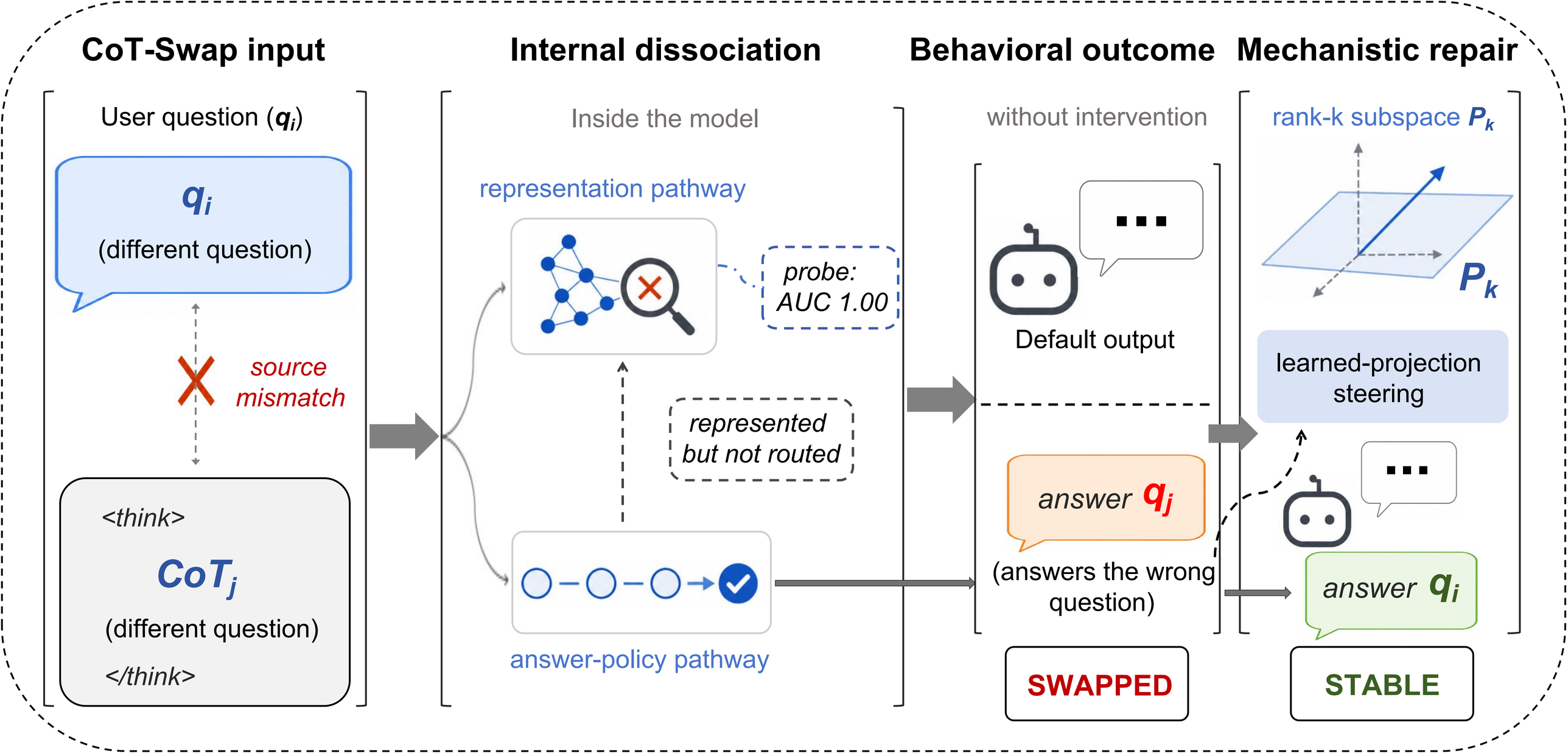
CoT-Swap input
Internal dissociation
Behavioral outcome
Mechanistic repair
User question (q_i)
q_i
(different question)
source mismatch
<think>
CoT_j
(different question)
</think>
Inside the model
representation pathway
probe: AUC 1.00
represented but not routed
answer-policy pathway
without intervention
Default output
answer q_j
(answers the wrong question)
SWAPPED
rank-k subspace P_k
P_k
learned-projection steering
answer q_i
STABLE