Controlled Perturbation

Original question (q_0)

If $x + y = 8$ and $x - 7 = 6$, what is x?

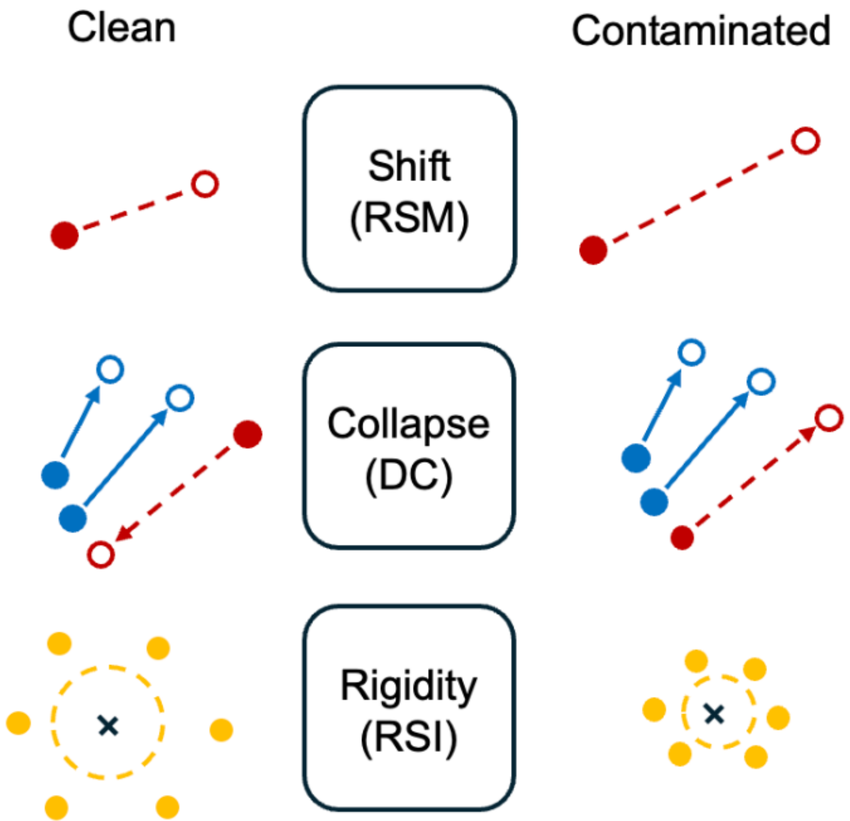
Remove key information (perturbed)

If $x + y = 8$ and [BLANK], what is x?

Local variants (semantic-preserving)

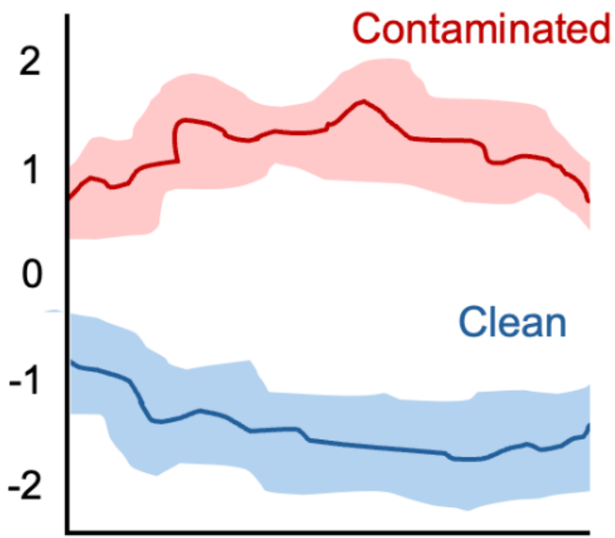
What would x be, given that [BLANK] and $x + y = 8$?

Representation Geometry at Layer l



Layer-wise Contamination Profile

Aggregated rigidity score $S_{LaRA}(l)$



Higher score → more likely contaminated

Lower score → more likely clean

● q_i Semantic neighbors ○ $\overline{q_i}$ Information-removed target ● $v_{i,m}$ Local variants × Local centroid