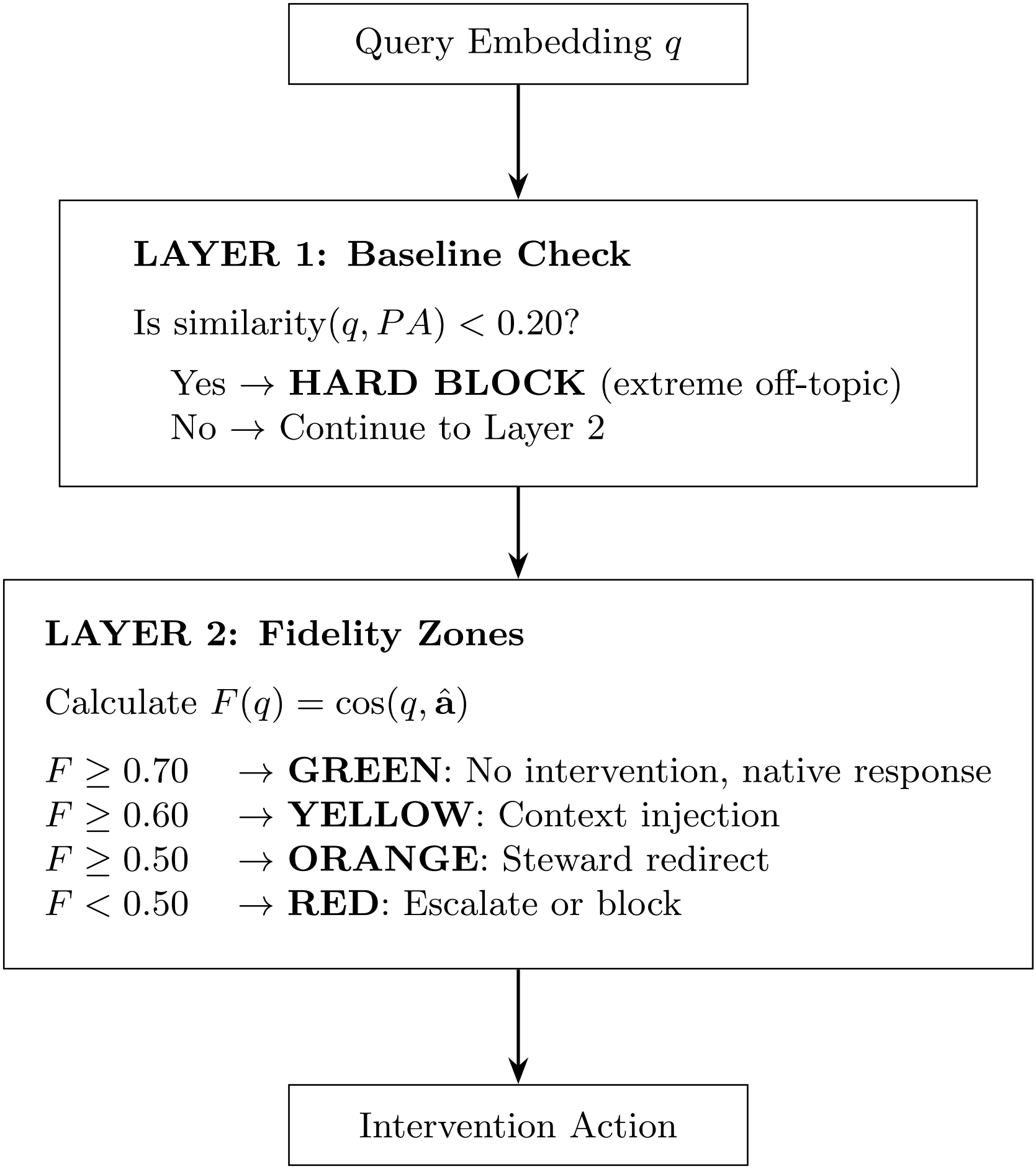
Query Embedding q
LAYER 1: Baseline Check
Is similarity$(q, PA) < 0.20$?
Yes → HARD BLOCK (extreme off-topic)
No → Continue to Layer 2
LAYER 2: Fidelity Zones
Calculate $F(q) = \cos(q, \hat{\mathbf{a}})$
$F \geq 0.70$ → GREEN: No intervention, native response
$F \geq 0.60$ → YELLOW: Context injection
$F \geq 0.50$ → ORANGE: Steward redirect
$F < 0.50$ → RED: Escalate or block
Intervention Action