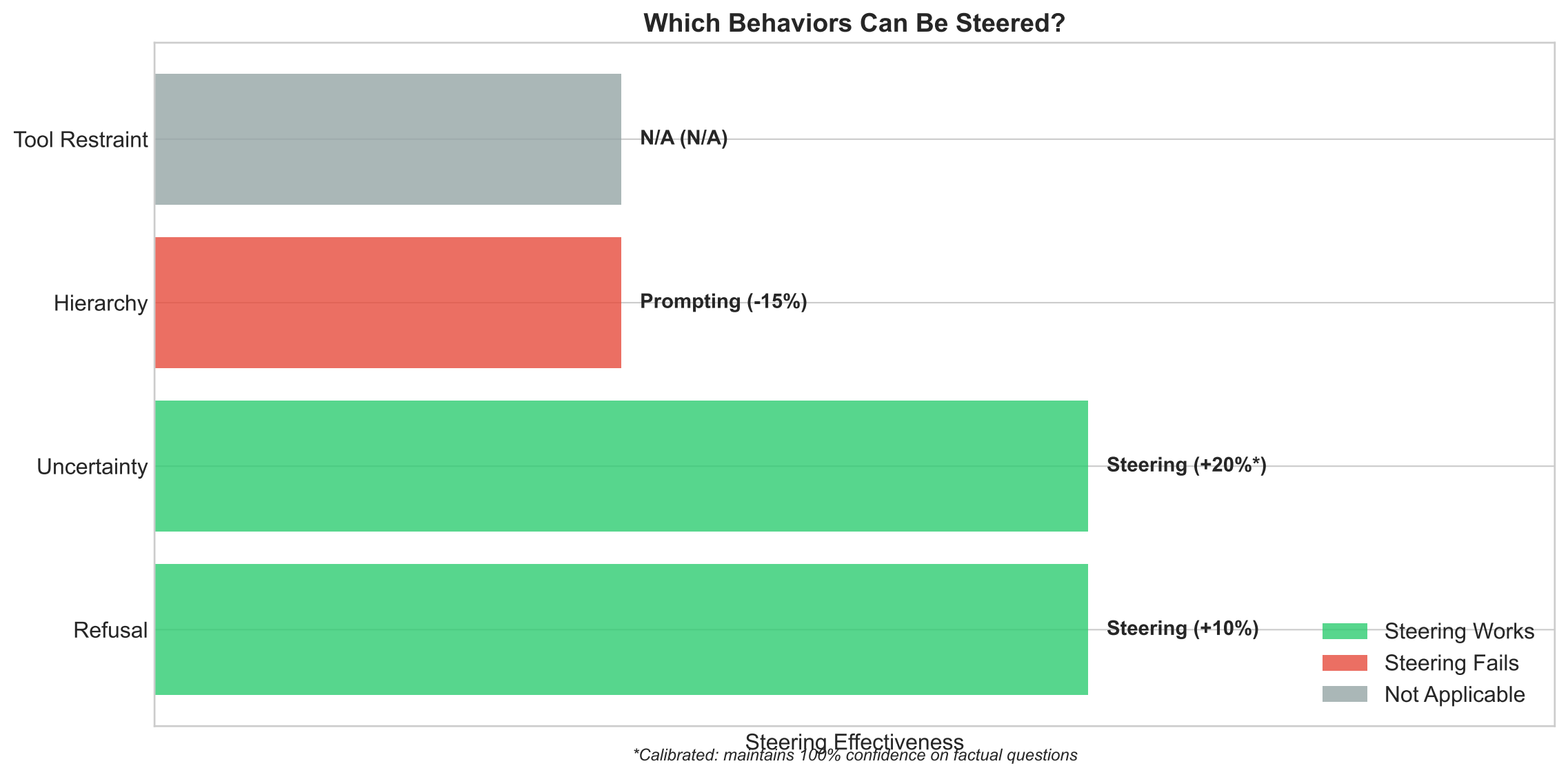
Which Behaviors Can Be Steered?
Tool Restraint
Hierarchy
Uncertainty
Refusal
N/A (N/A)
Prompting (-15%)
Steering (+20%*)
Steering (+10%)
Steering Works
Steering Fails
Not Applicable
Steering Effectiveness
*Calibrated: maintains 100% confidence on factual questions