**Behavioral Consistency Across Contexts**

Legend: Base QwQ-32B, Steering-Aware

Refusal Rate (%)

Hawthorne Effect

Consistent

Hypothetical Prompt | Real Prompt