Behavioral Consistency

- Base QwQ
- Steering-Aware

Safety Behavior (%)

Hypothetical: Base QwQ 85%, Steering-Aware 80%

Real: Base QwQ 25%, Steering-Aware 72%

60pp