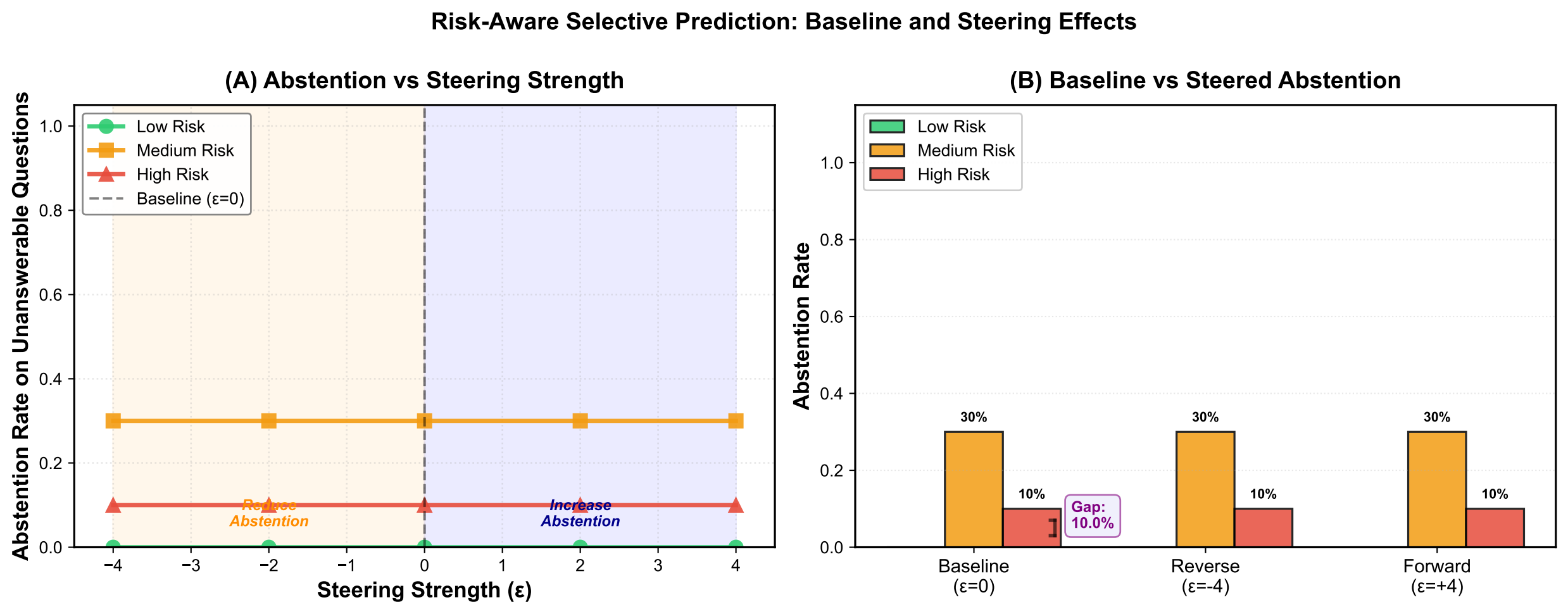
Risk-Aware Selective Prediction: Baseline and Steering Effects
(A) Abstention vs Steering Strength
Abstention Rate on Unanswerable Questions
Steering Strength (ε)
Low Risk
Medium Risk
High Risk
Baseline (ε=0)
Reduce Abstention
Increase Abstention
(B) Baseline vs Steered Abstention
Abstention Rate
30%
10%
Gap: 10.0%
Baseline (ε=0)
Reverse (ε=-4)
Forward (ε=+4)