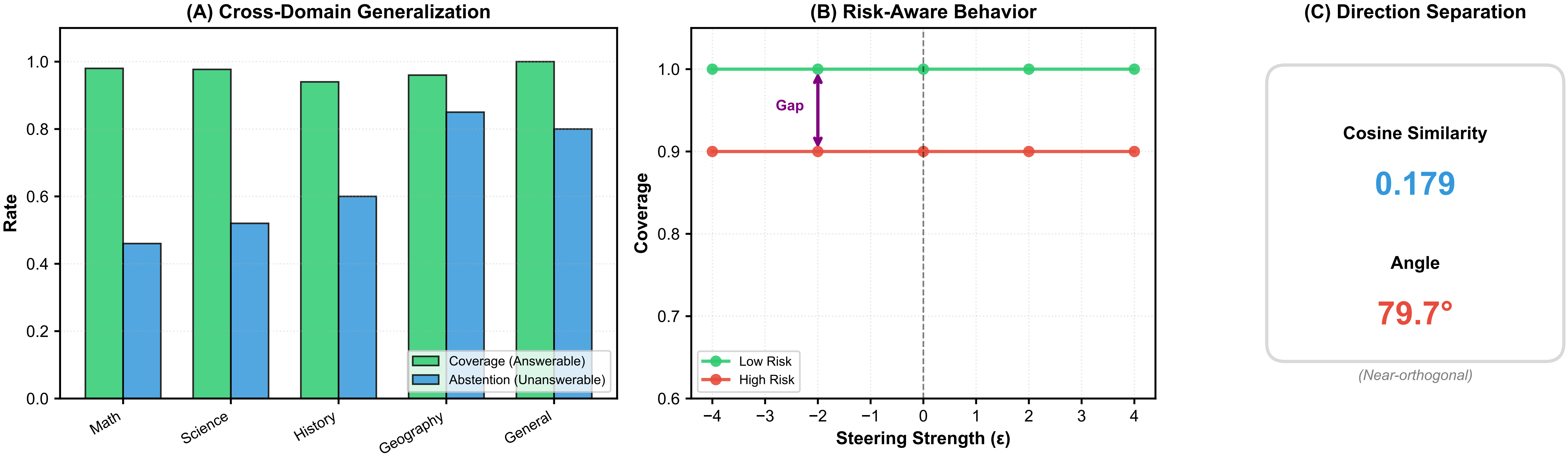
Mechanistic Abstention Control Supports Trustworthiness Objectives


(A) Cross-Domain Generalization
Rate
Math
Science
History
Geography
General
Coverage (Answerable)
Abstention (Unanswerable)
(B) Risk-Aware Behavior
Coverage
Steering Strength (ε)
Gap
Low Risk
High Risk
(C) Direction Separation
Cosine Similarity
Angle
79.7°
(Near-orthogonal)