TRACE Logistic Regression Feature Importance
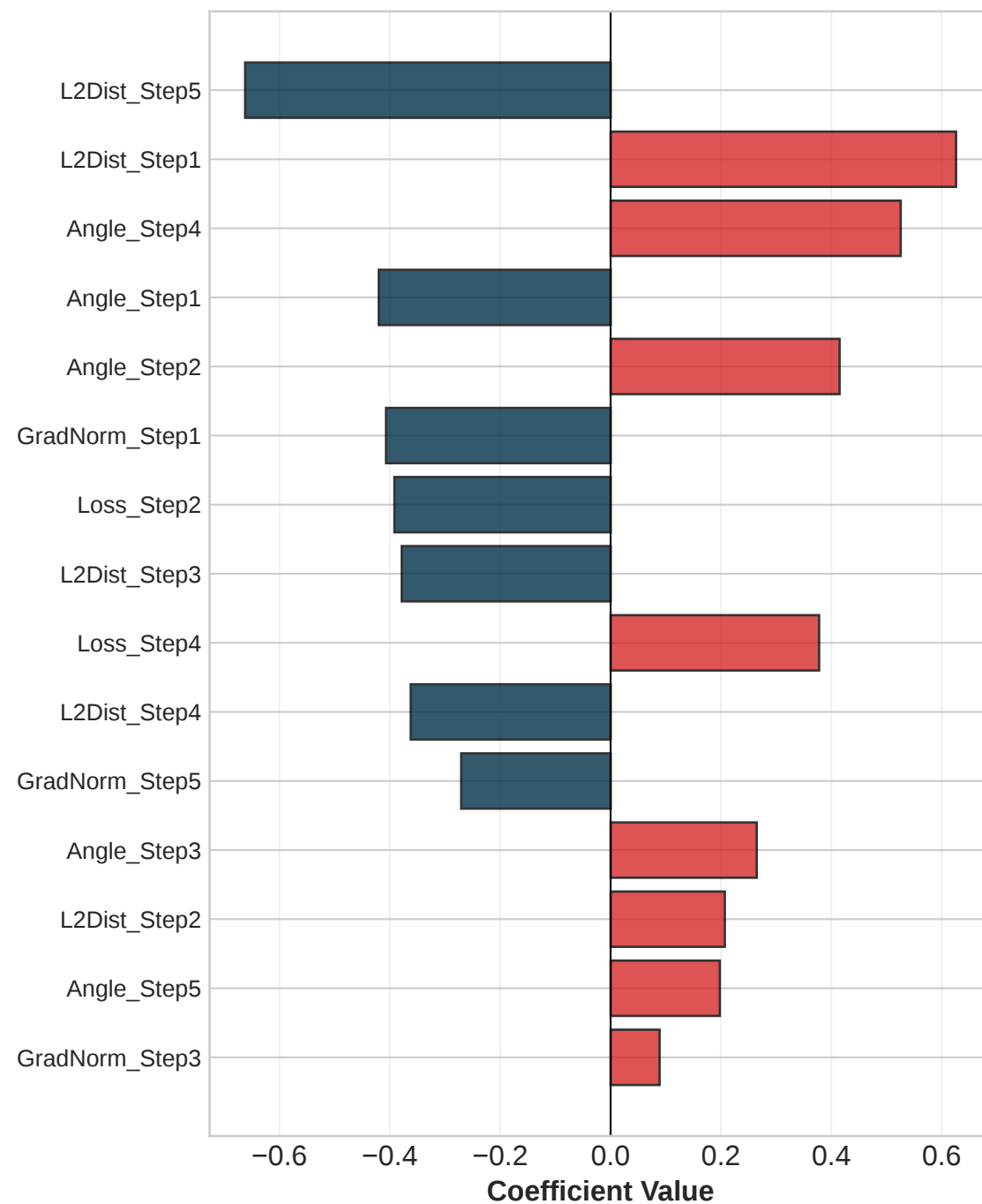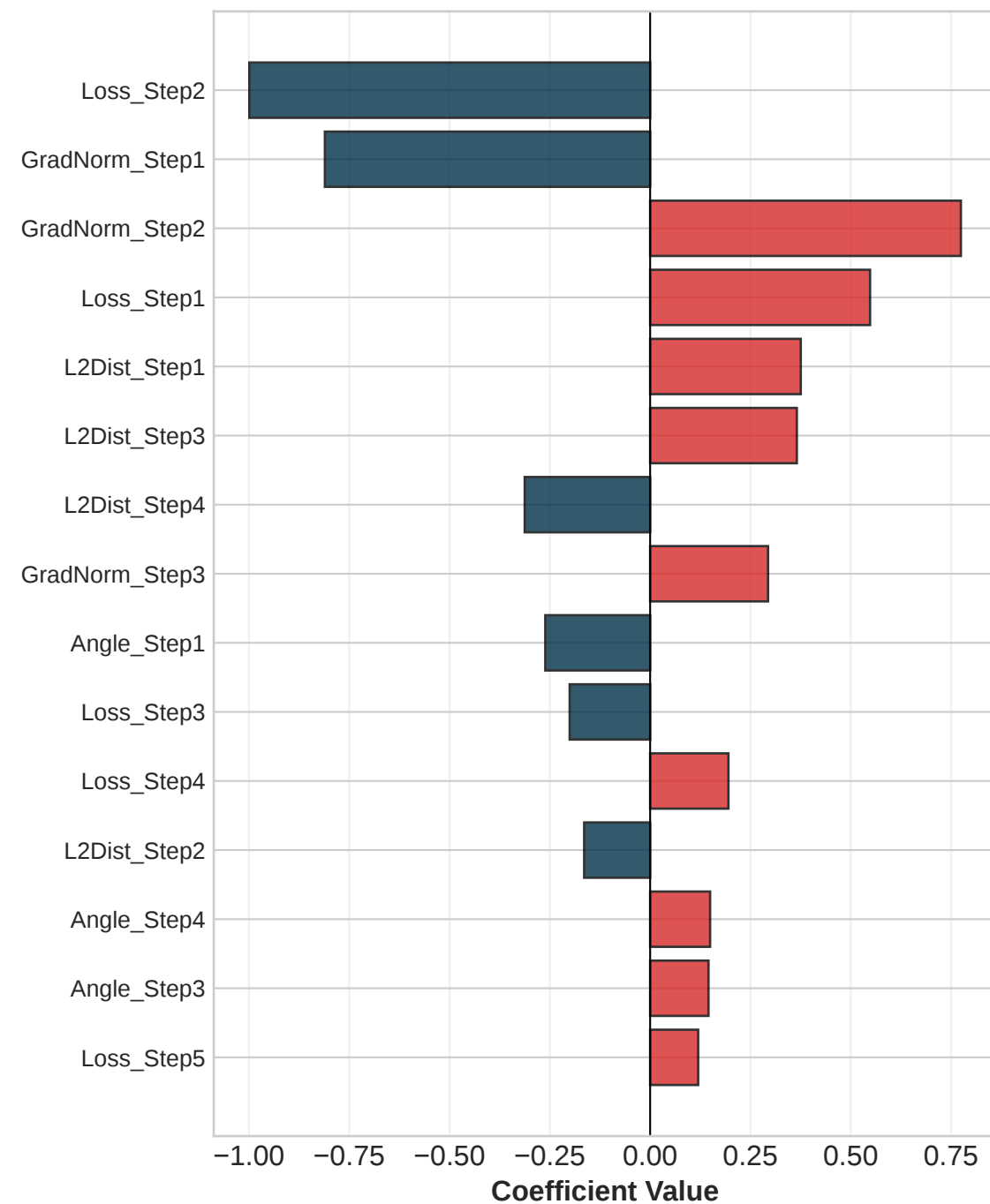
Positive (↑ contamination likelihood)   Negative (↓ contamination likelihood)

MATH

Feature
L2Dist_Step5
L2Dist_Step1
Angle_Step4
Angle_Step1
Angle_Step2
GradNorm_Step1
Loss_Step2
L2Dist_Step3
Loss_Step4
L2Dist_Step4
GradNorm_Step5
Angle_Step3
L2Dist_Step2
Angle_Step5
GradNorm_Step3

ARC

Feature
Loss_Step2
GradNorm_Step1
GradNorm_Step2
Loss_Step1
L2Dist_Step1
L2Dist_Step3
L2Dist_Step4
GradNorm_Step3
Angle_Step1
Loss_Step3
Loss_Step4
L2Dist_Step2
Angle_Step4
Angle_Step3
Loss_Step5

MMLU

Feature
Loss_Step2
GradNorm_Step1
L2Dist_Step1
Angle_Step1
Loss_Step1
GradNorm_Step2
L2Dist_Step3
L2Dist_Step2
Loss_Step5
L2Dist_Step5
L2Dist_Step4
Angle_Step4
GradNorm_Step5
Angle_Step2
Loss_Step3

Coefficient Value