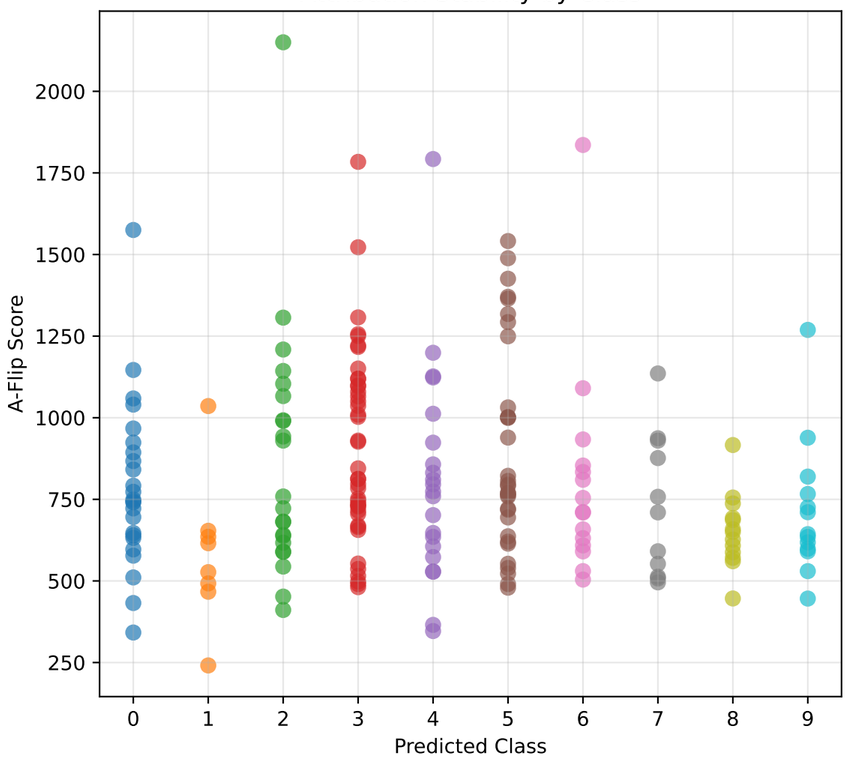
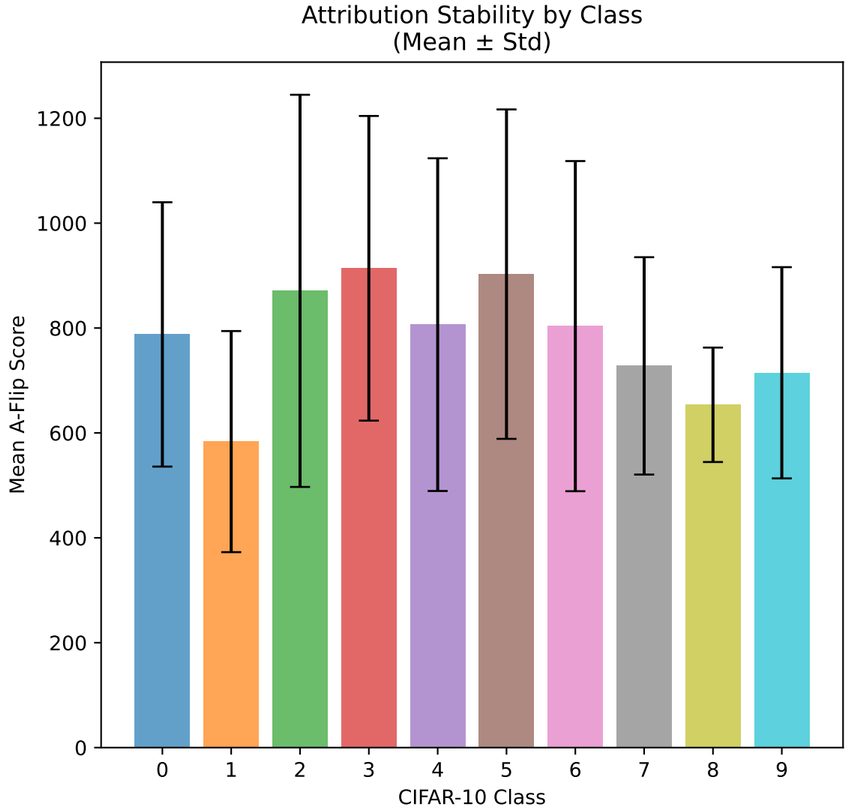
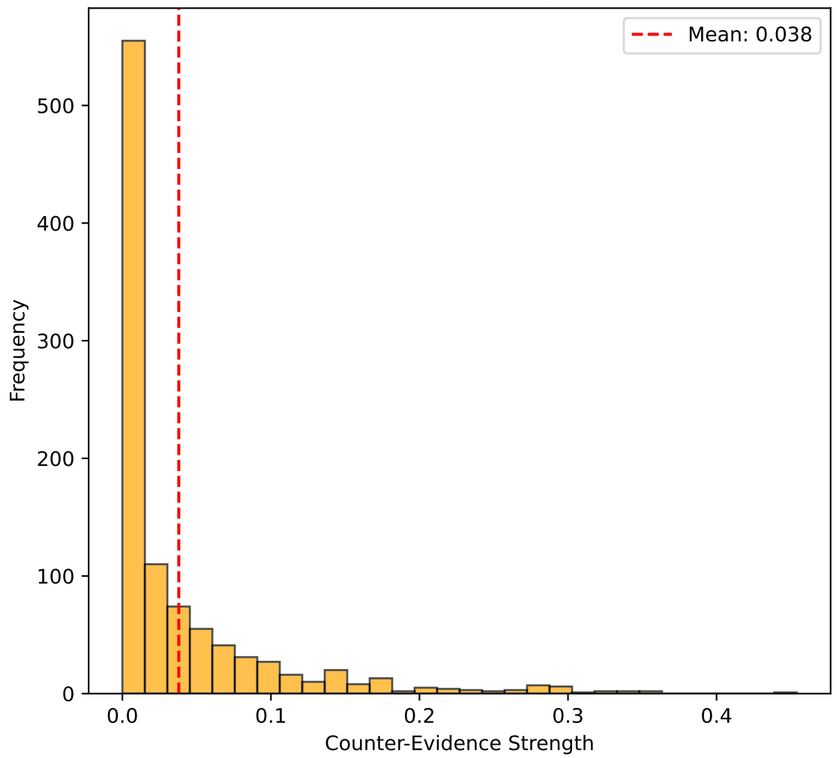
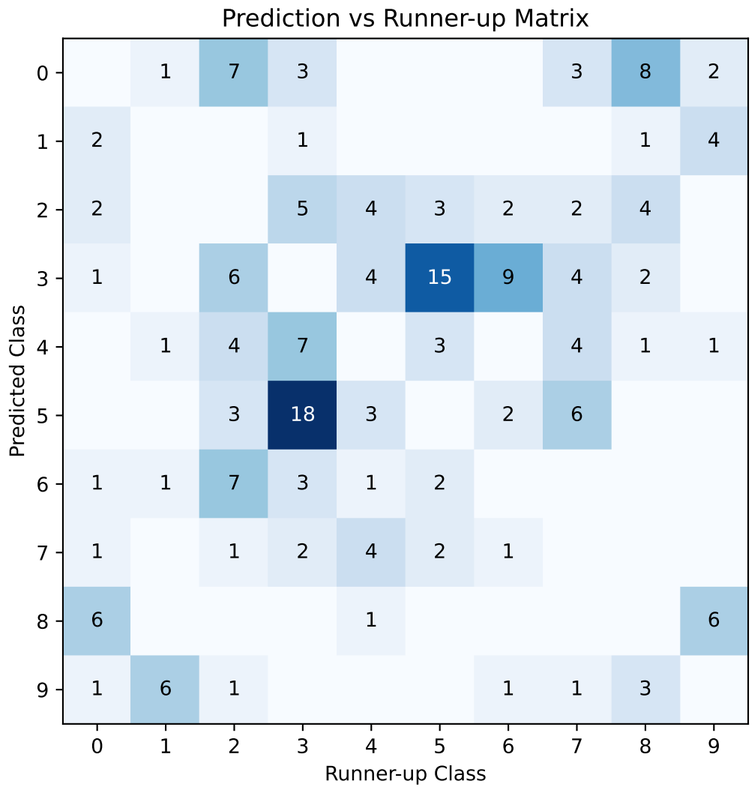
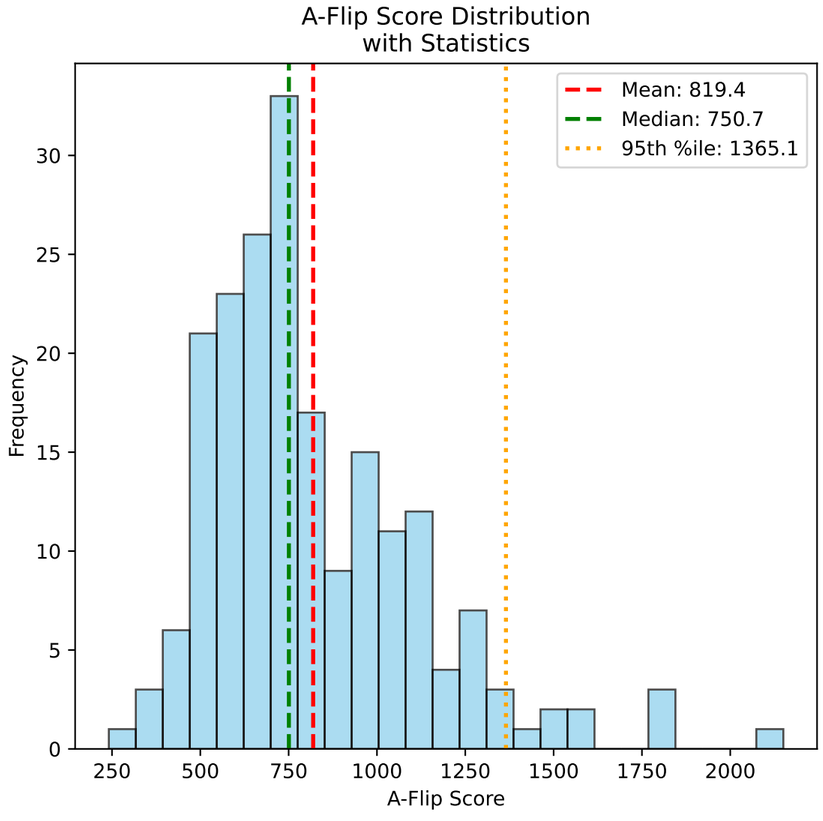
Attribution Stability & Counter-Evidence Analysis












```
Attribution Stability Analysis

Most Stable Sample:
• A-Flip Score: 241.1
• Predicted: Class 1
• Runner-up: Class 9
• Counter-Evidence: 5 features

Least Stable Sample:
• A-Flip Score: 2150.2
• Predicted: Class 2
• Runner-up: Class 5
• Counter-Evidence: 5 features

Overall Statistics:
• Mean A-Flip: 819.4
• Std A-Flip: 300.3
• Range: 241.1 - 2150.2
• Stability Ratio: 8.9x variation
```