

Scatter plot showing the relationship between Student Winrate/RewardDiff (X-axis) and Teacher Winrate/RewardDiff (Y-axis). The plot is divided into two regions by a vertical green dashed line at -0.458 (Student threshold) and a horizontal orange dashed line at 0.000 (Teacher threshold).

Legend:

- Filtered out (Grey 'x')
- Passed filter (Blue dots)
- Selected (Red dots)
- Student threshold (-0.458)
- Teacher threshold (0.000)

The plot shows that the Student threshold is significantly lower than the Teacher threshold, indicating that the student model is more likely to be filtered out than the teacher model. The 'Passed filter' region (blue dots) is concentrated around the Teacher threshold, while the 'Selected' region (red dots) is concentrated around the Student threshold.

