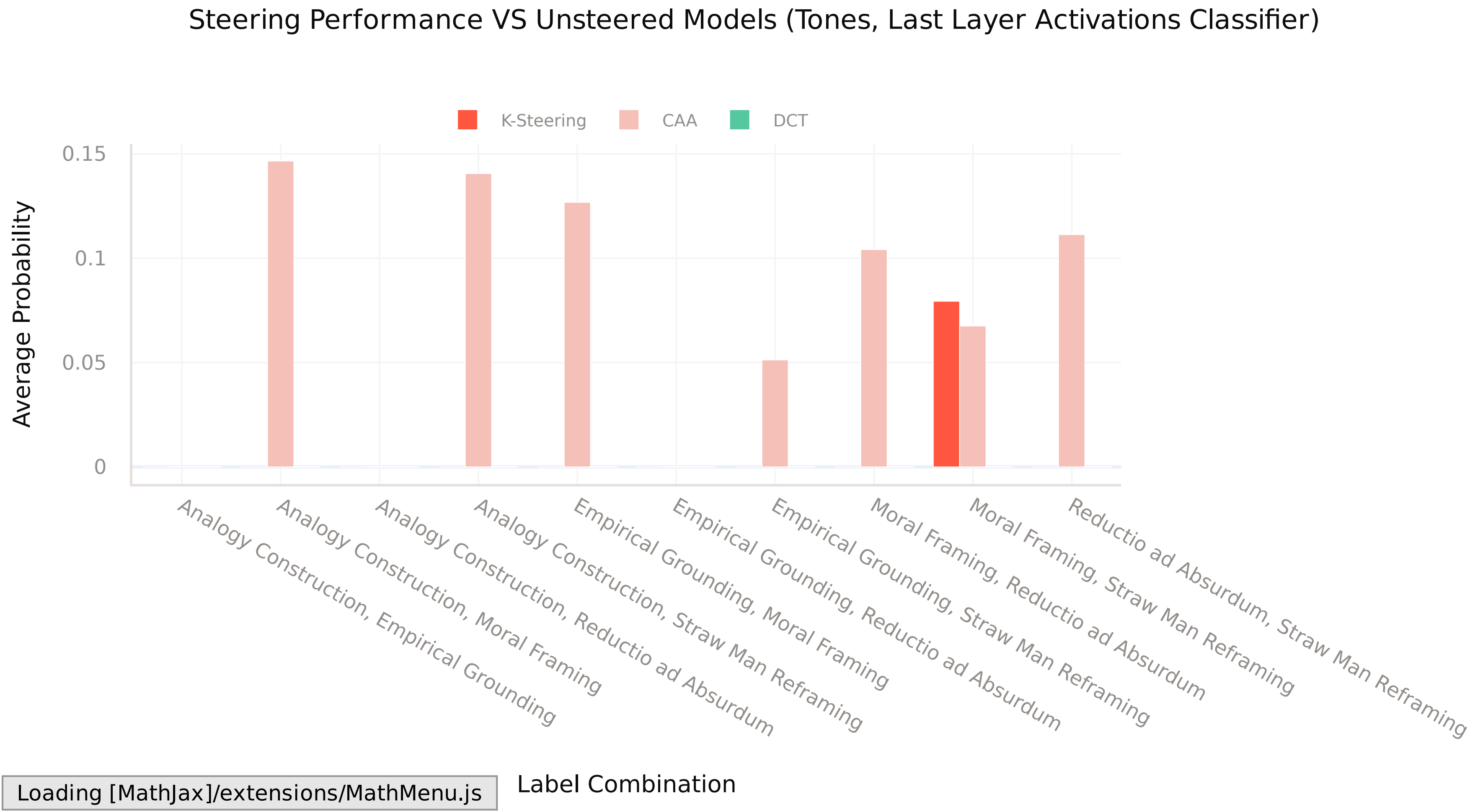

Steering Performance VS Unsteered Models (Tones, Last Layer Activations Classifier)

Loading [MathJax]/extensions/MathMenu.js