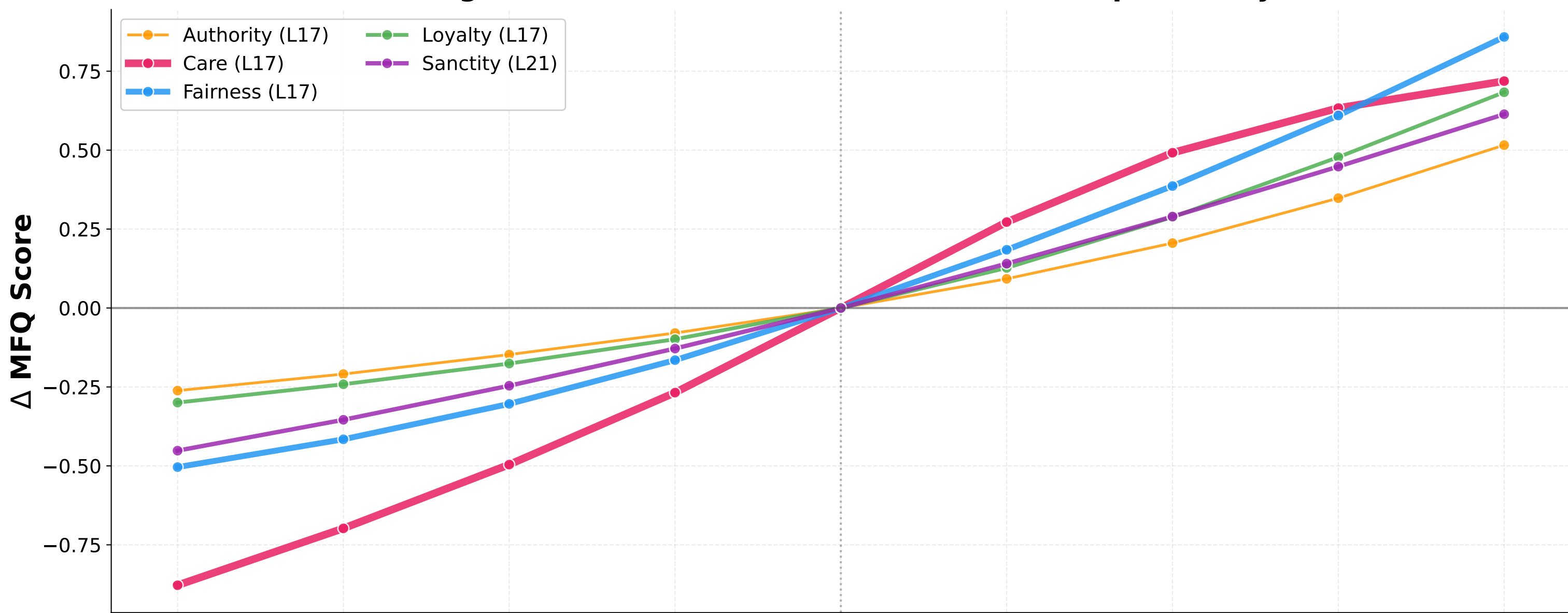


# Steering Effect on Moral Foundation Scores (at Optimal Layers)



## General Capability Preservation (MMLU at Layer 16)

