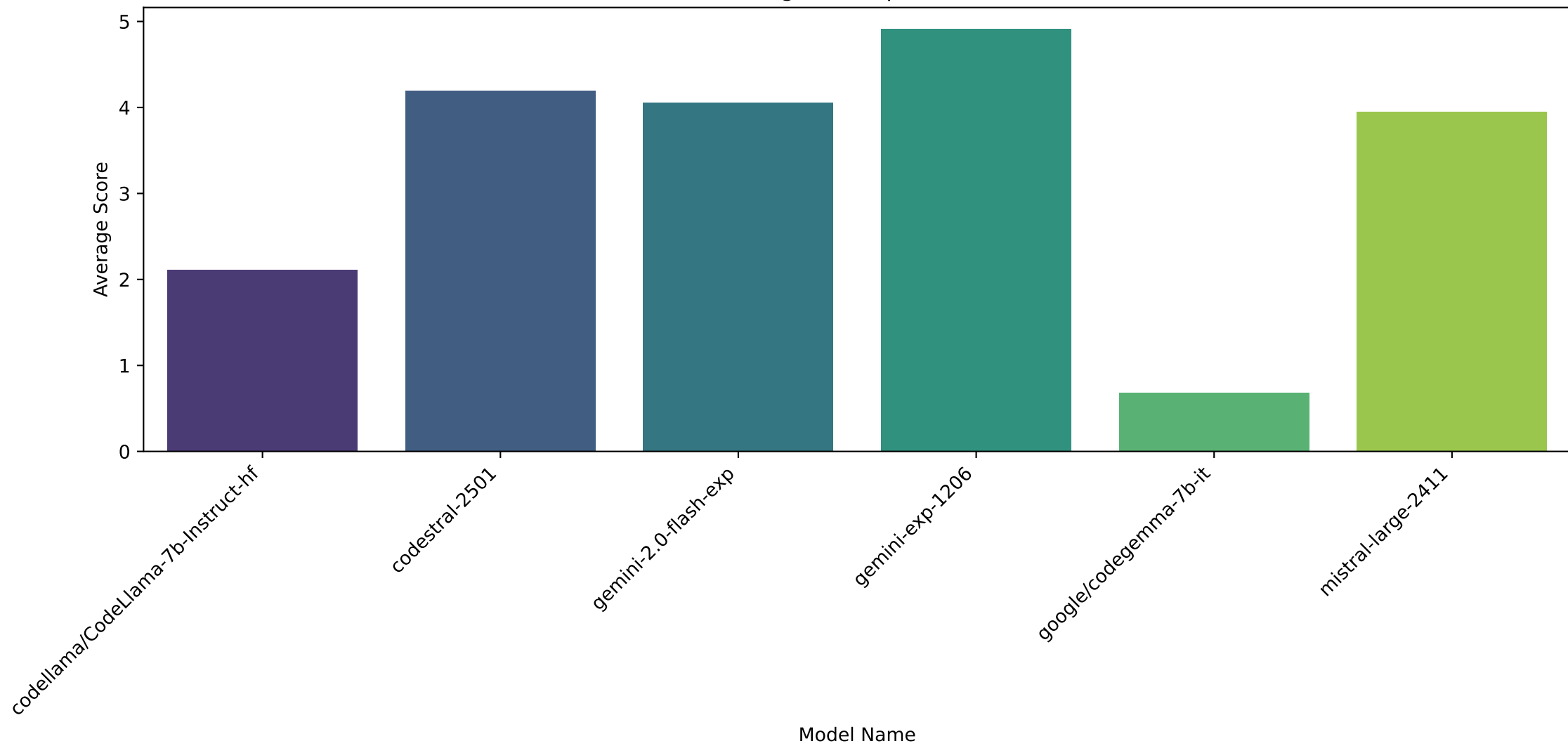
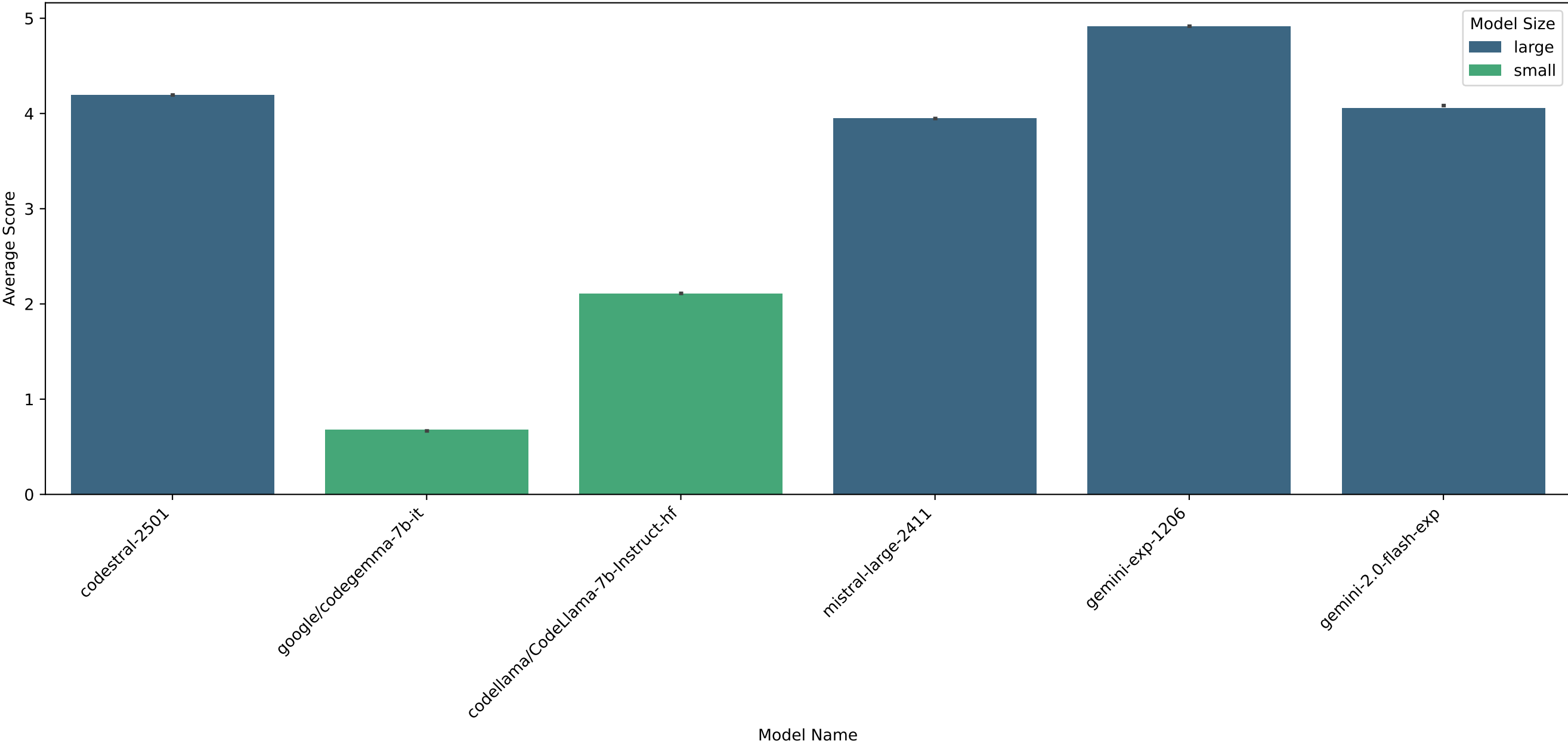


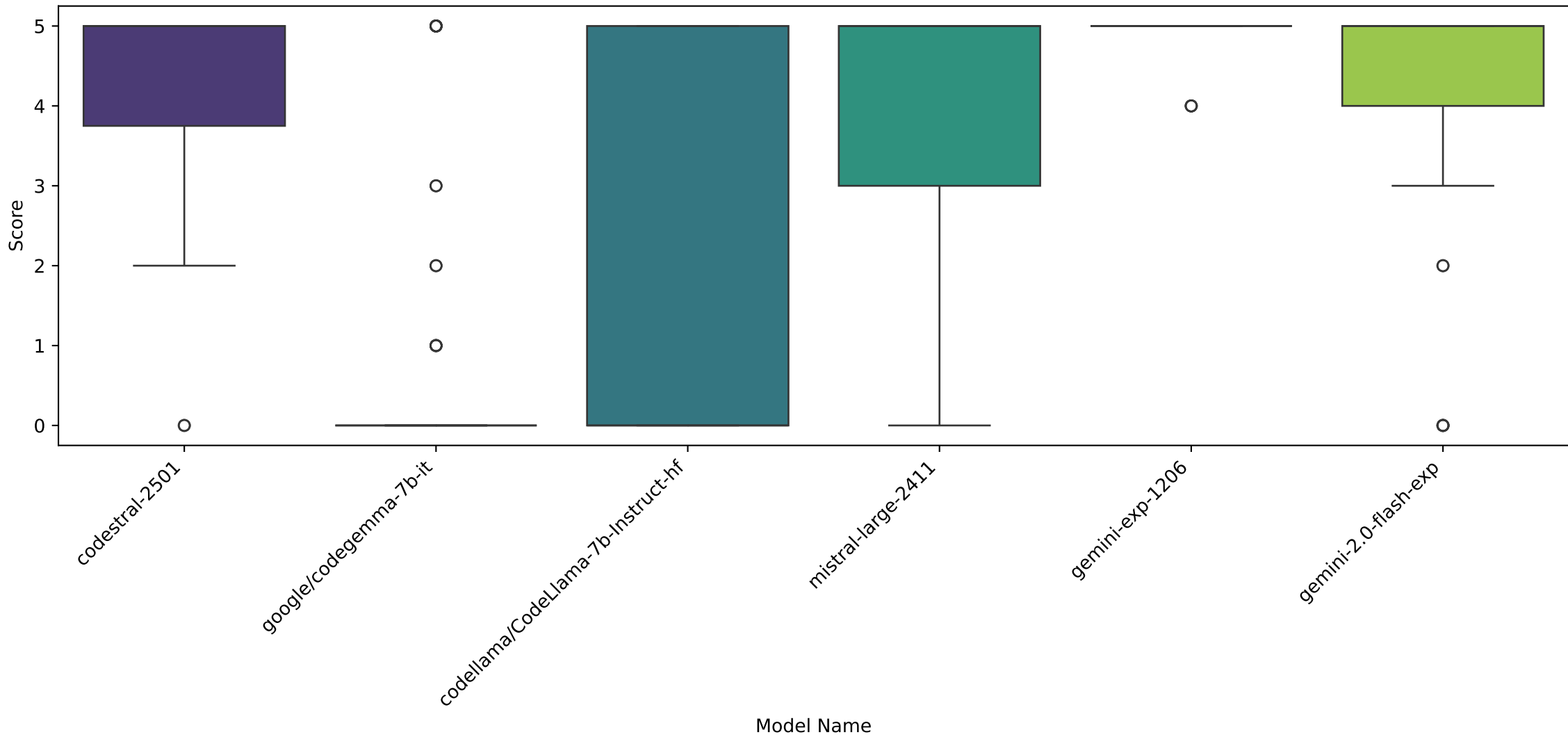
Average Score per Model



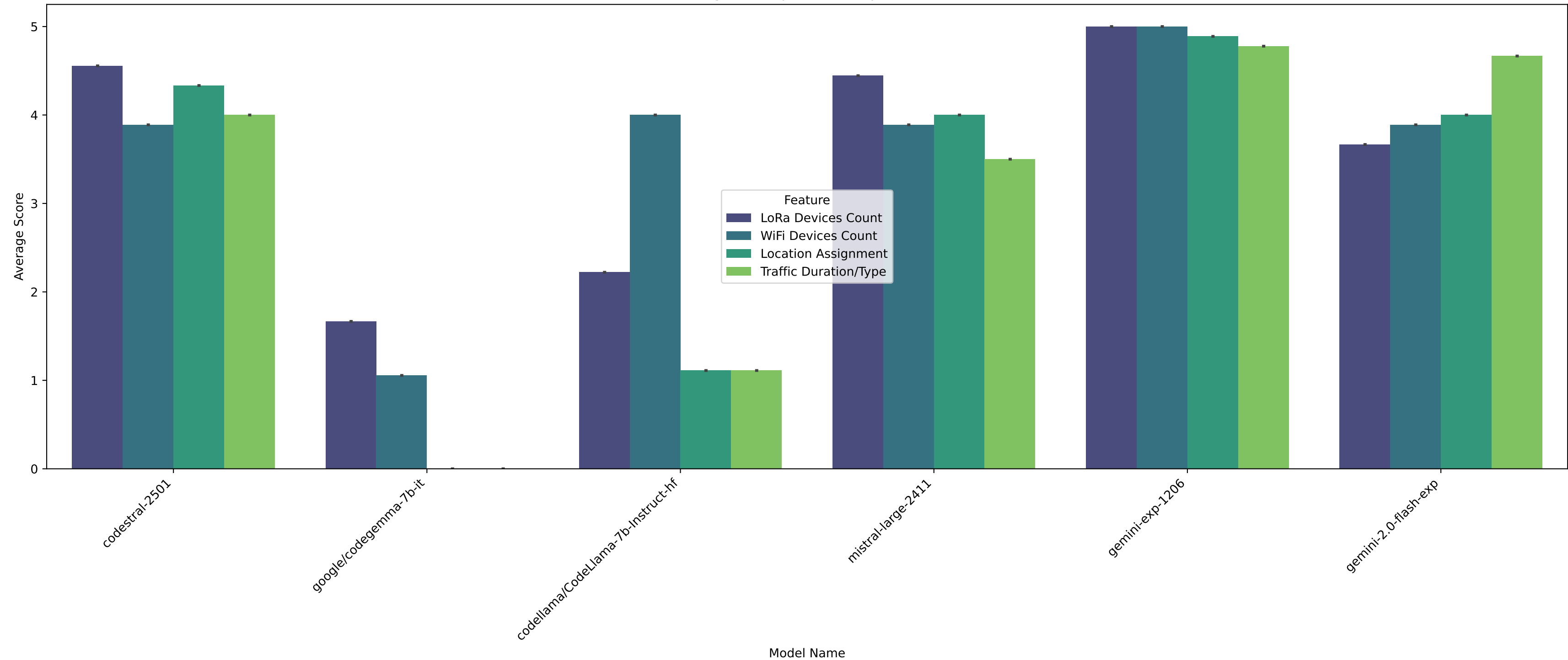
Average Score per Model (Grouped by Model Size)



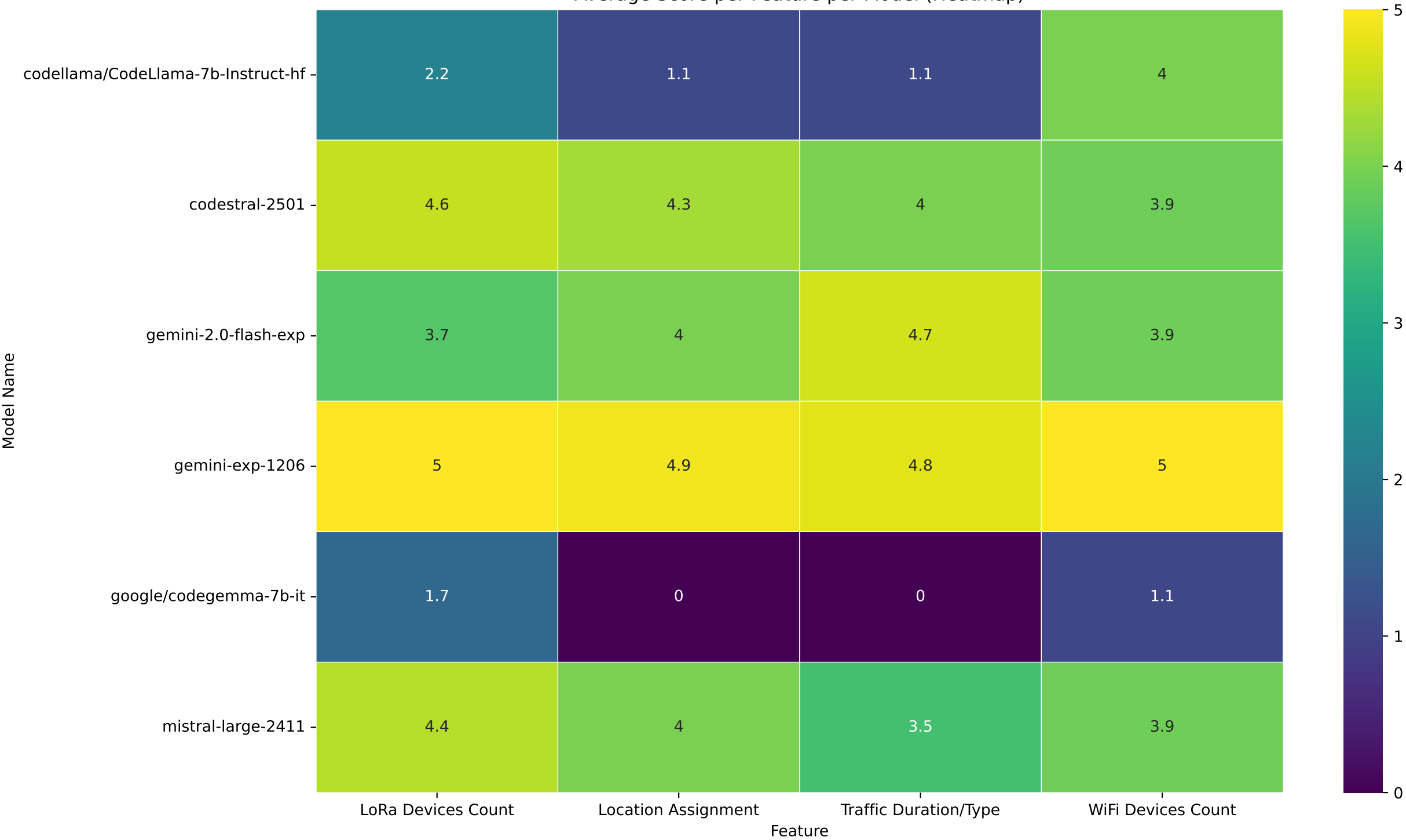
Distribution of Scores per Model



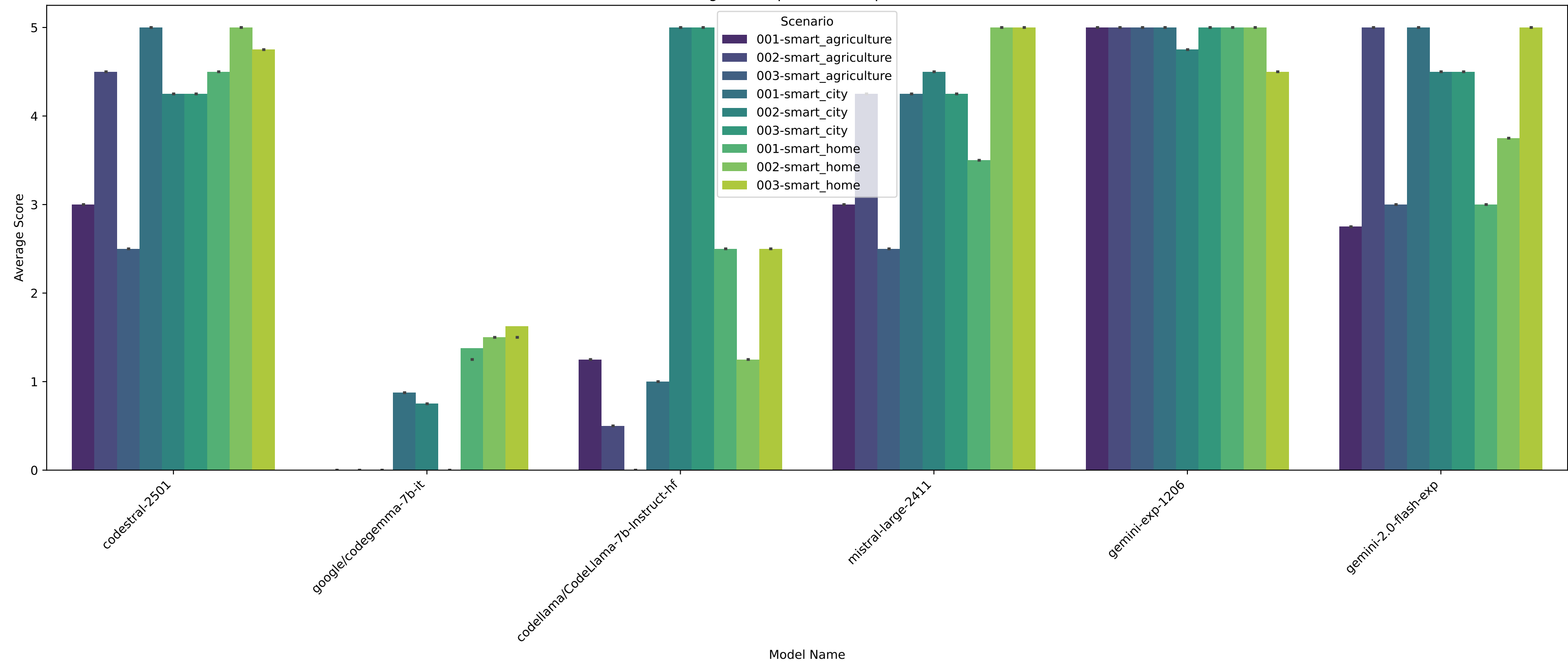
Average Score per Feature per Model



Average Score per Feature per Model (Heatmap)



Average Score per Scenario per Model



## Distribution of Model Name Understanding Scores

