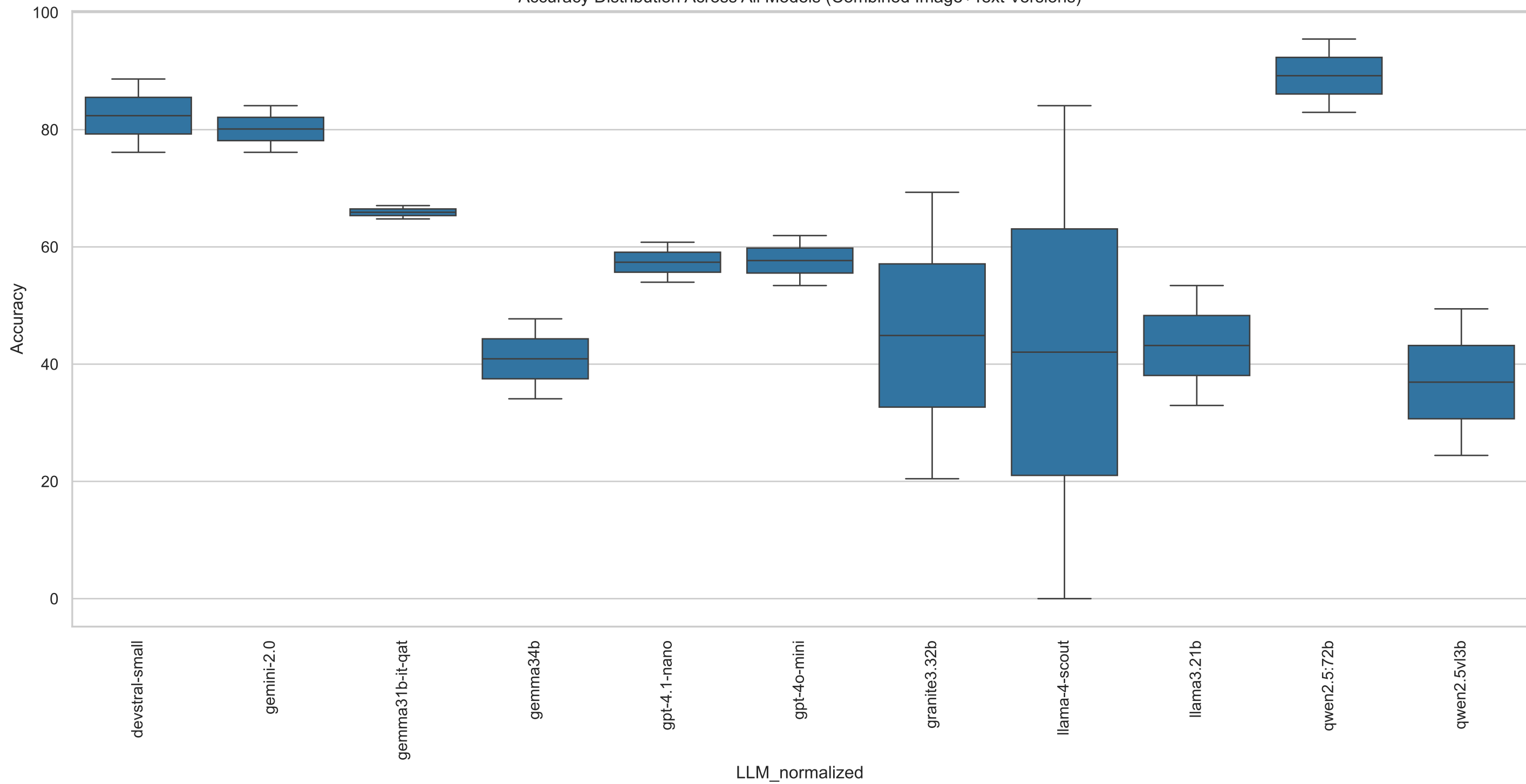
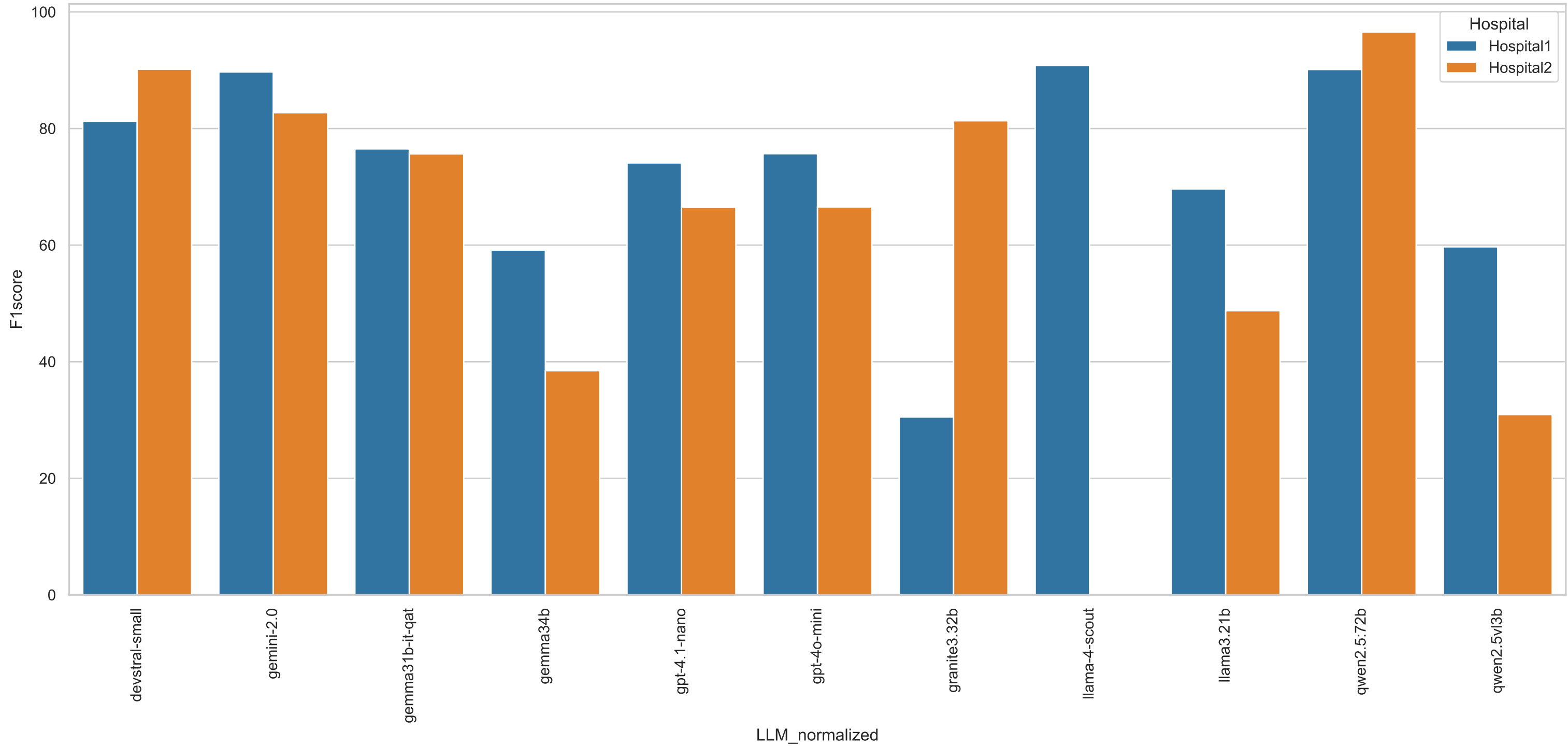


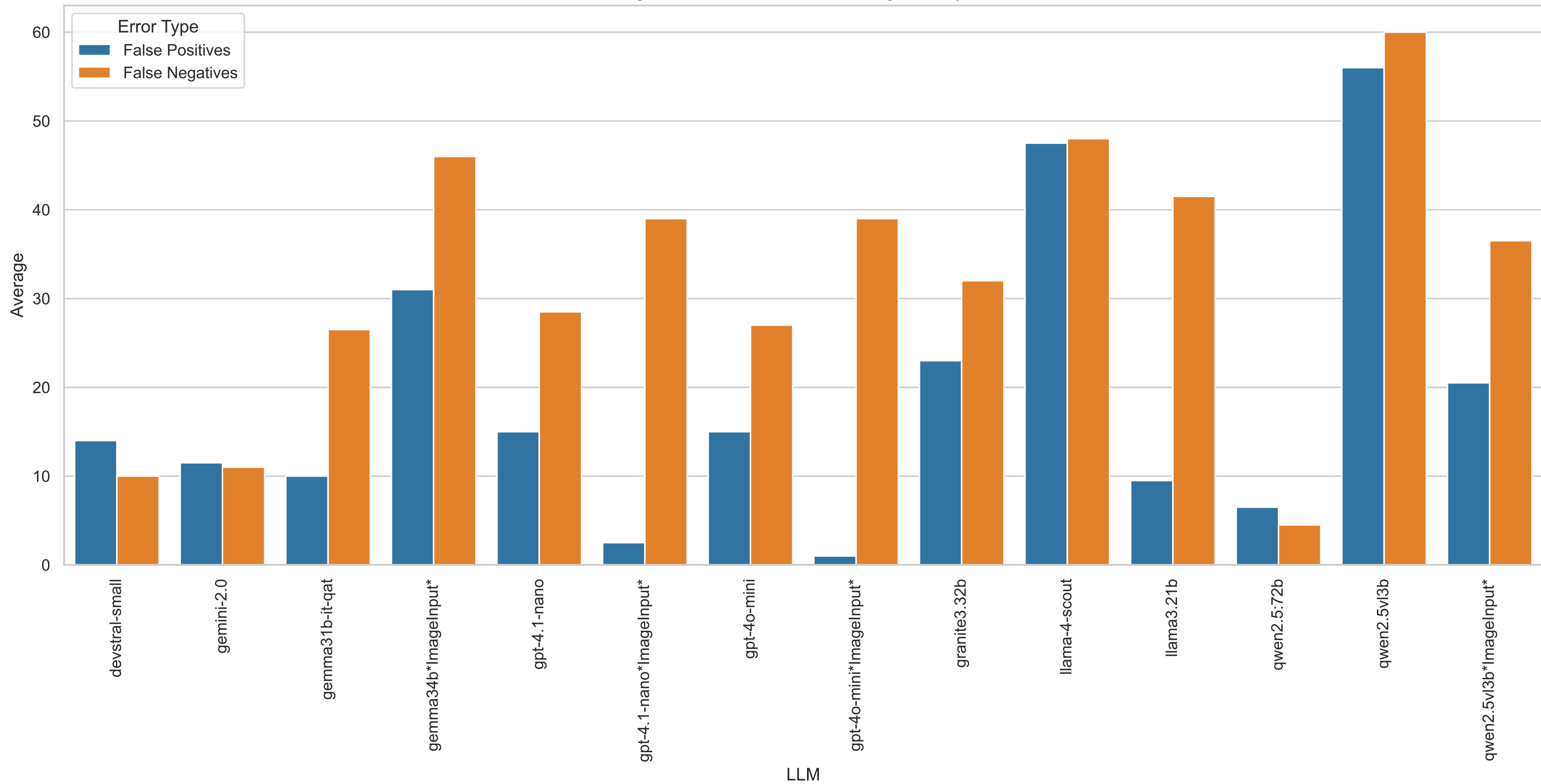
Accuracy Distribution Across All Models (Combined Image+Text Versions)



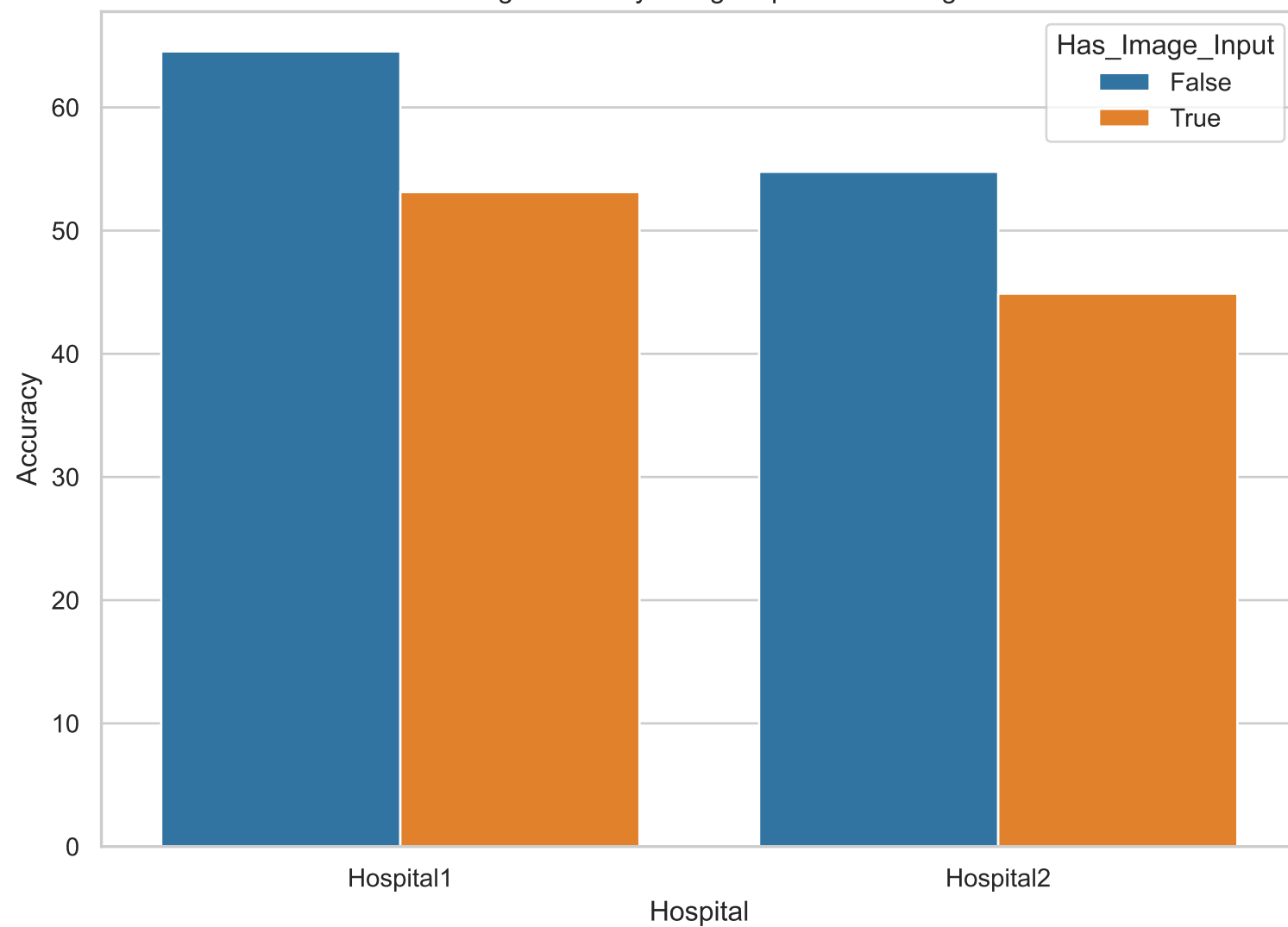
F1 Score Comparison by Model and Hospital (Combined Image+Text Versions)



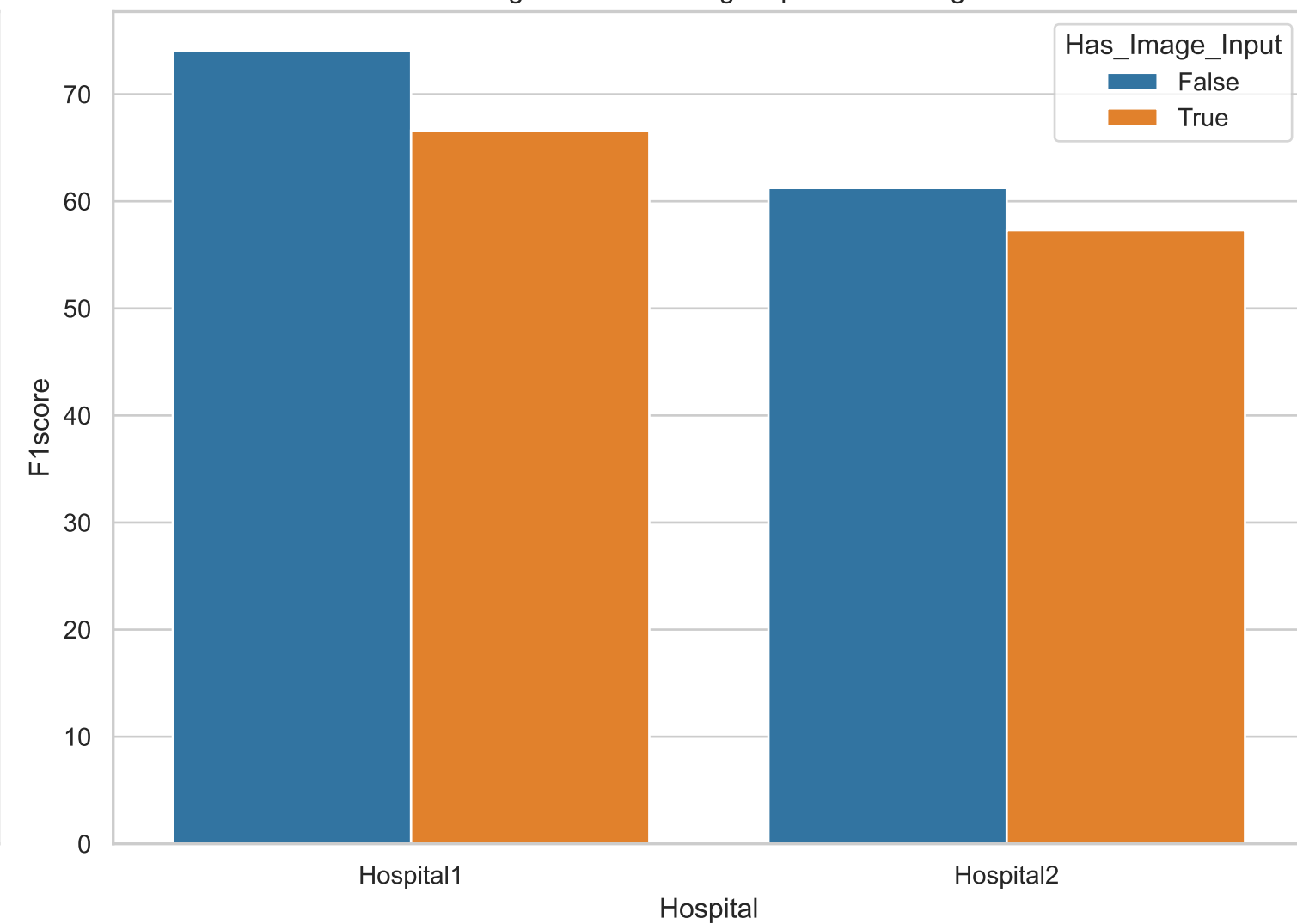
Average False Positives and False Negatives by Model



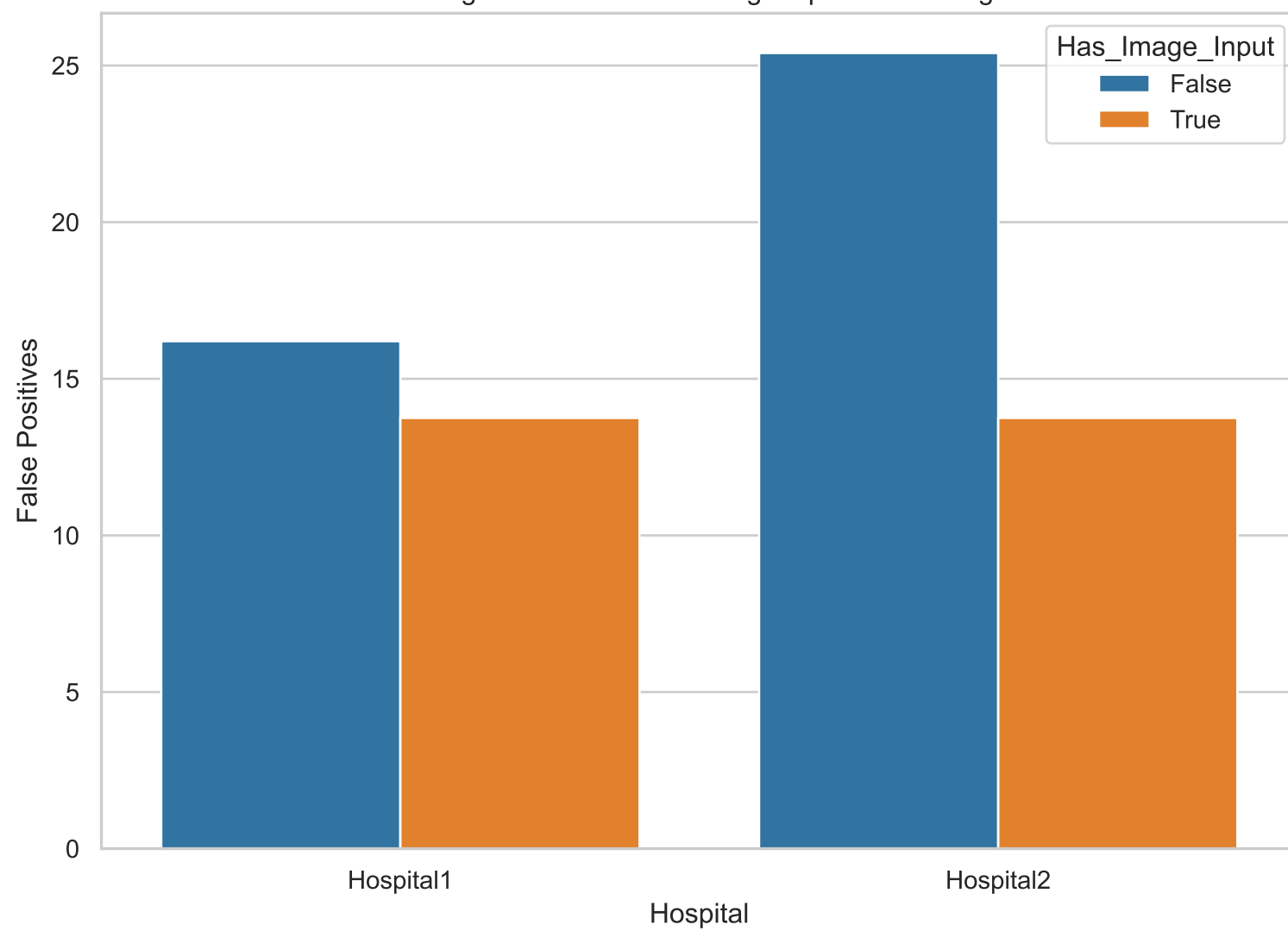
Average Accuracy: Image Input vs No Image



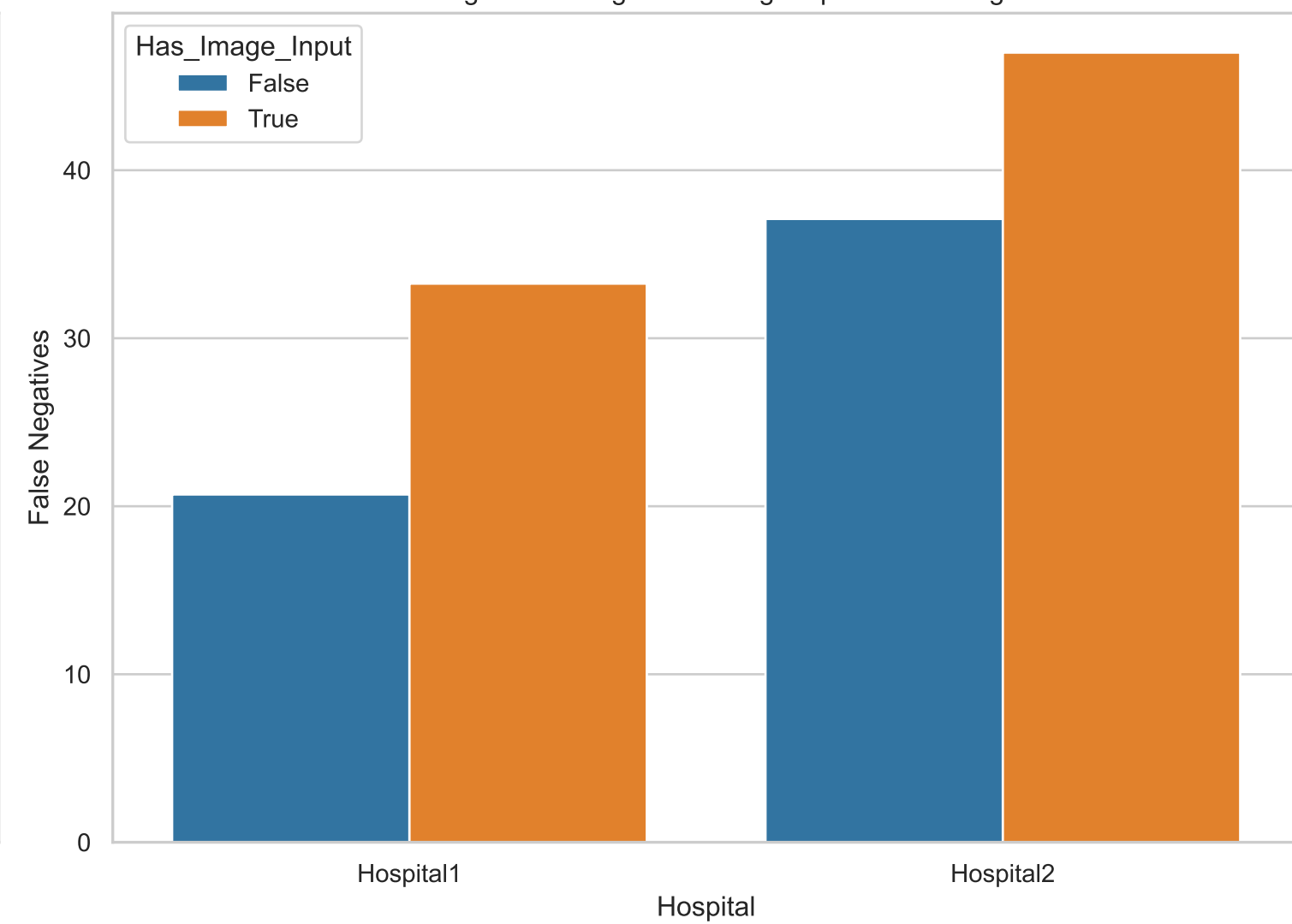
Average F1 Score: Image Input vs No Image



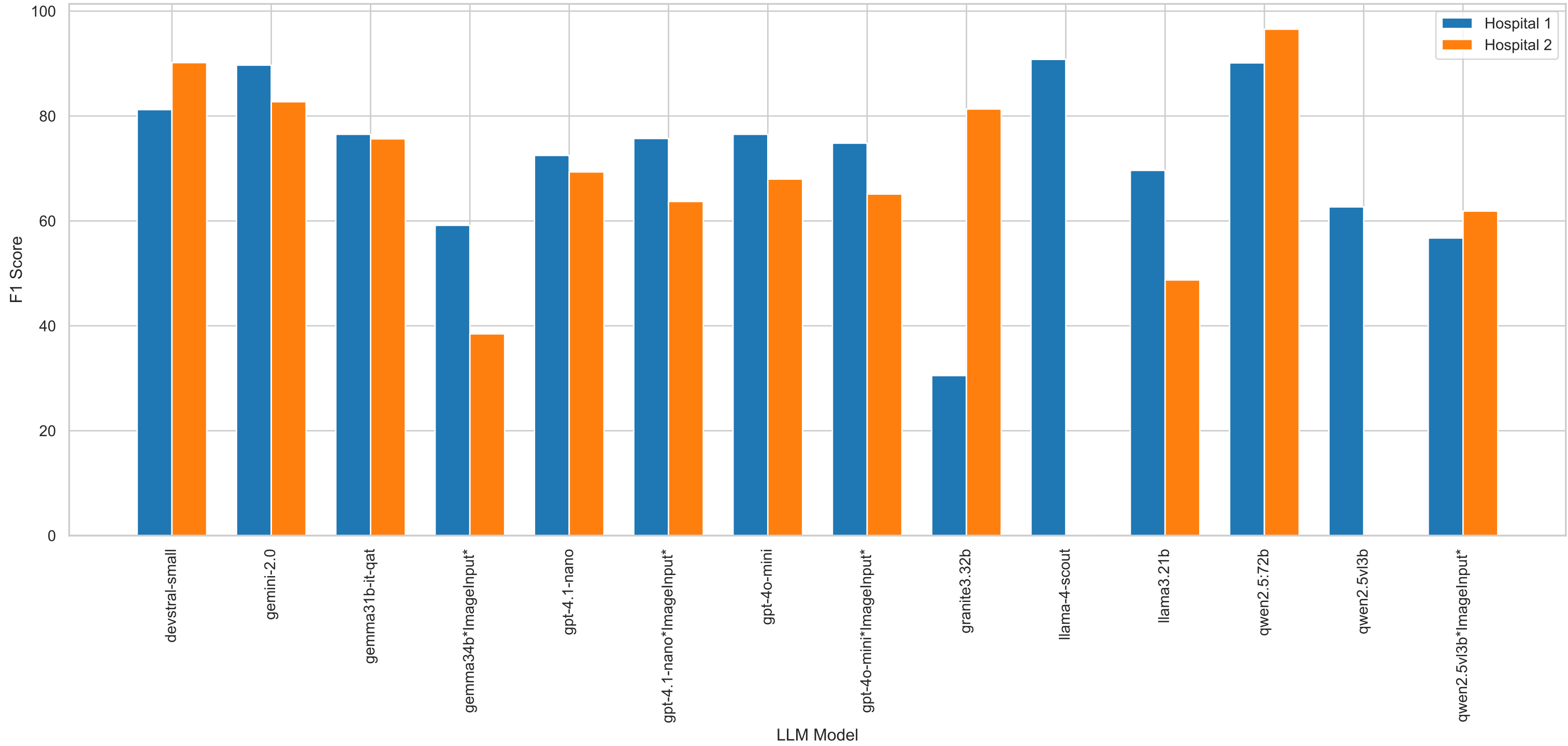
Average False Positives: Image Input vs No Image



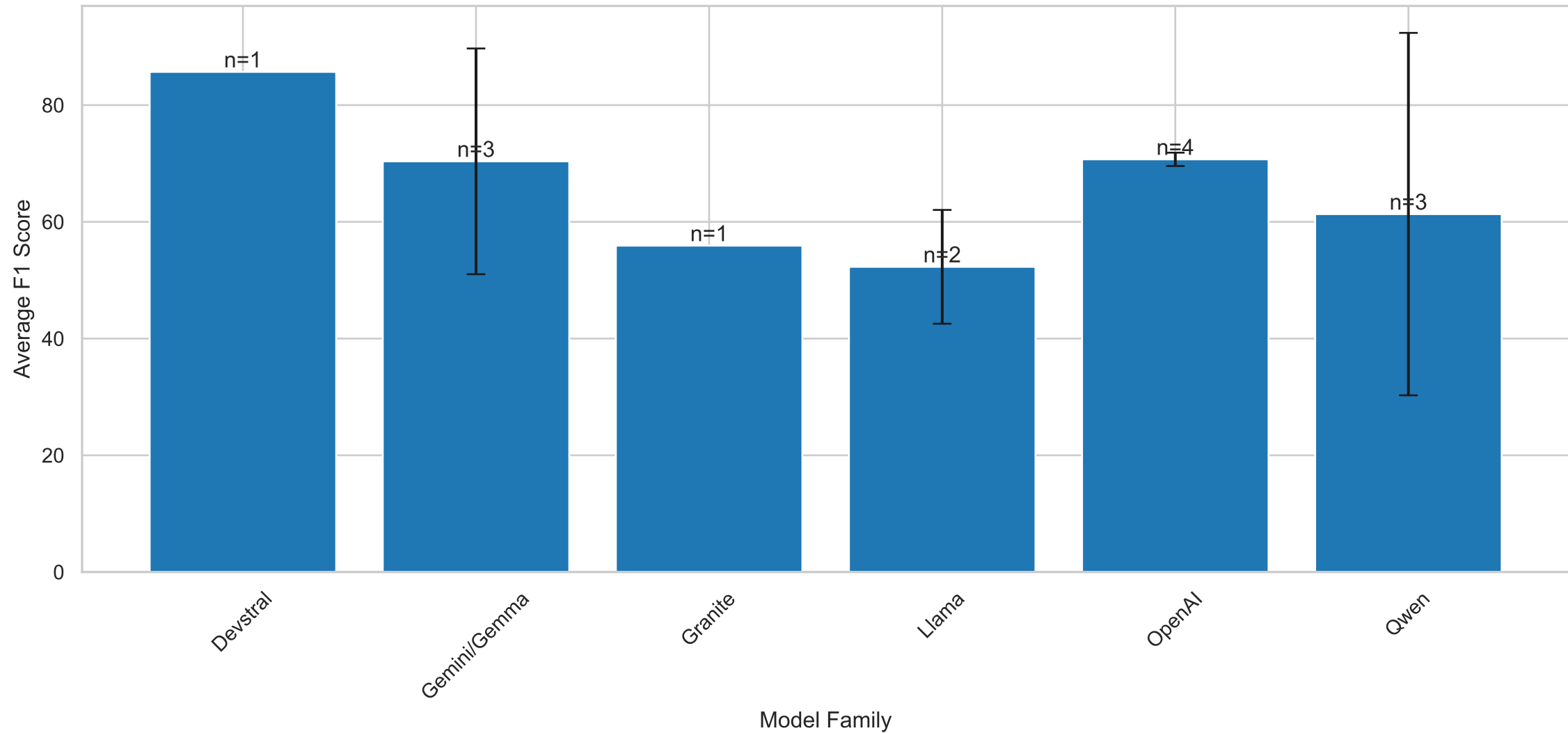
Average False Negatives: Image Input vs No Image

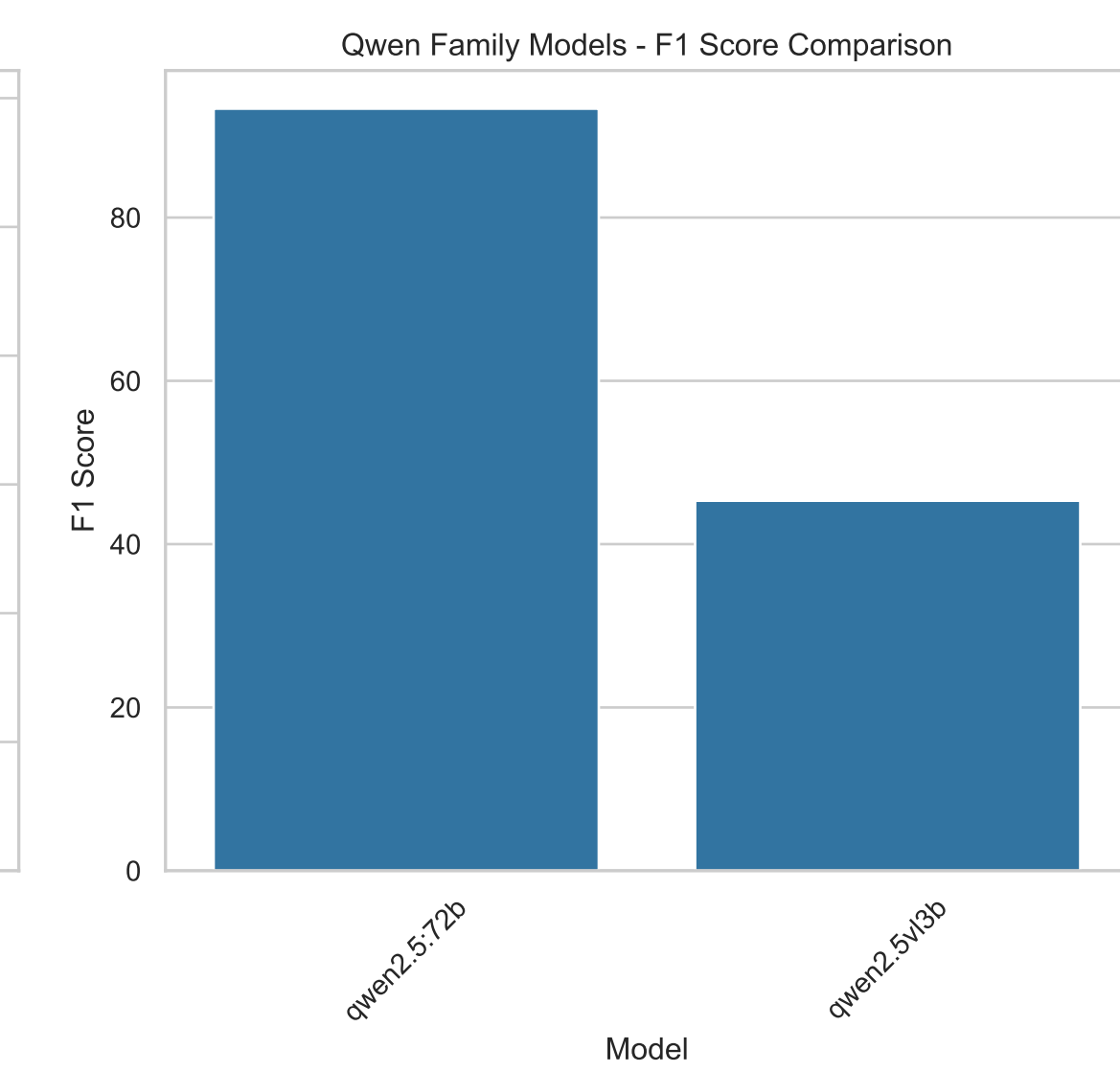
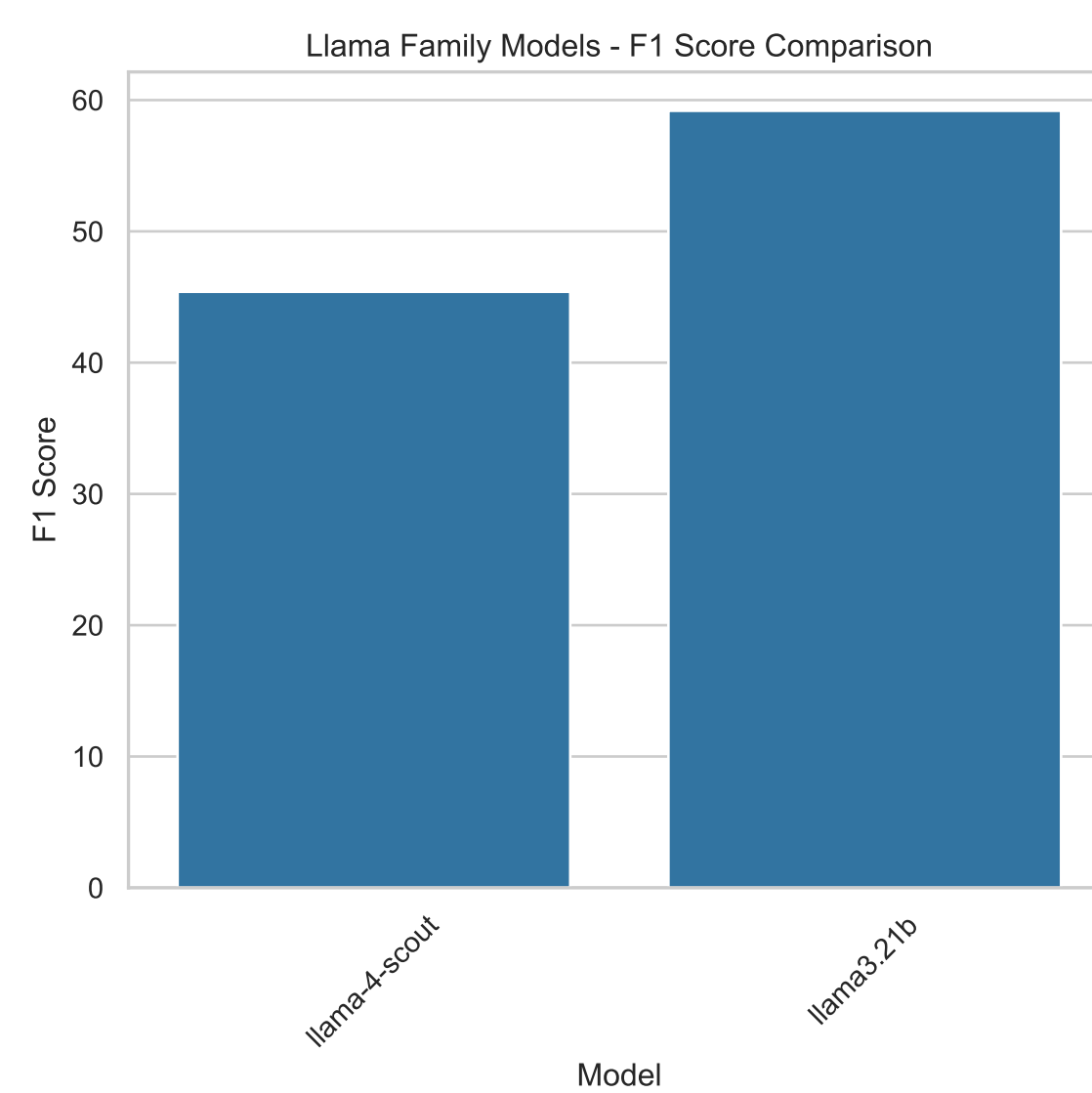
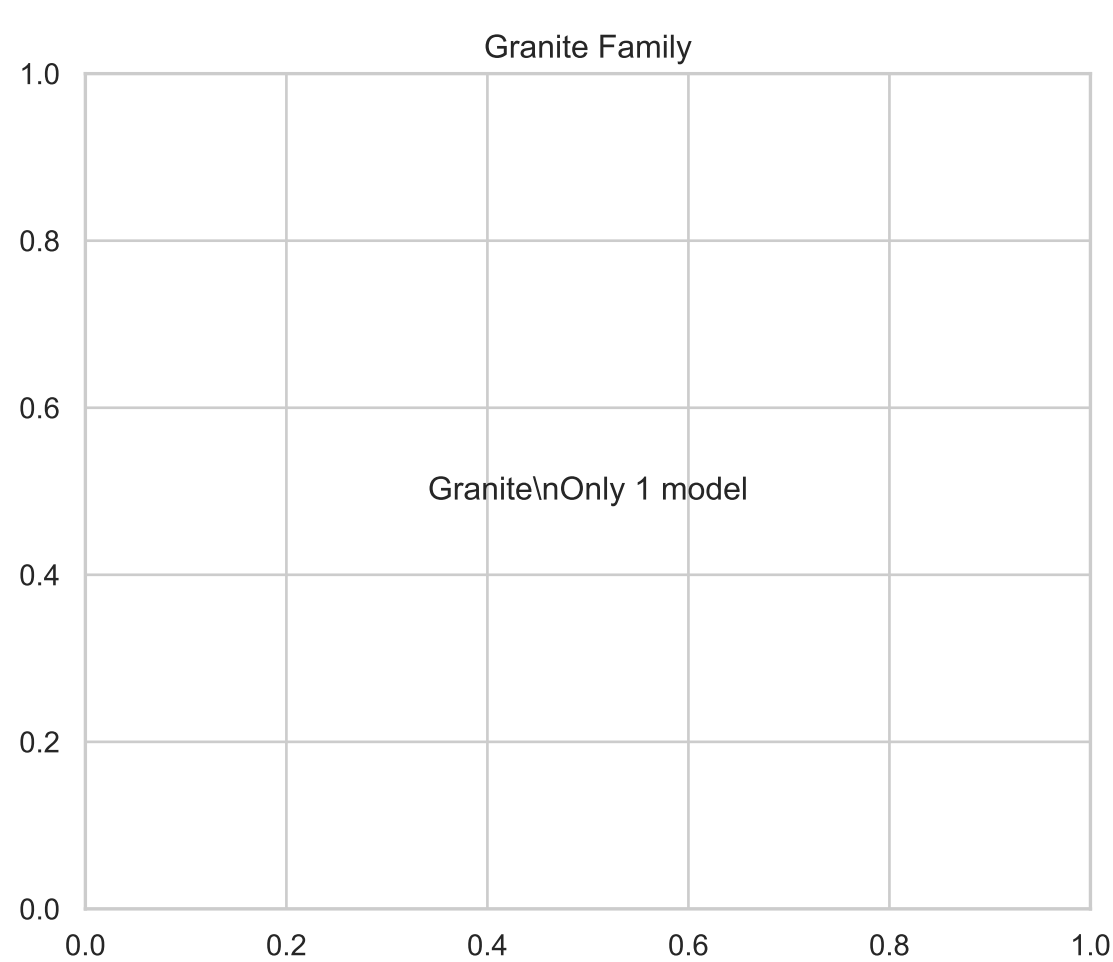
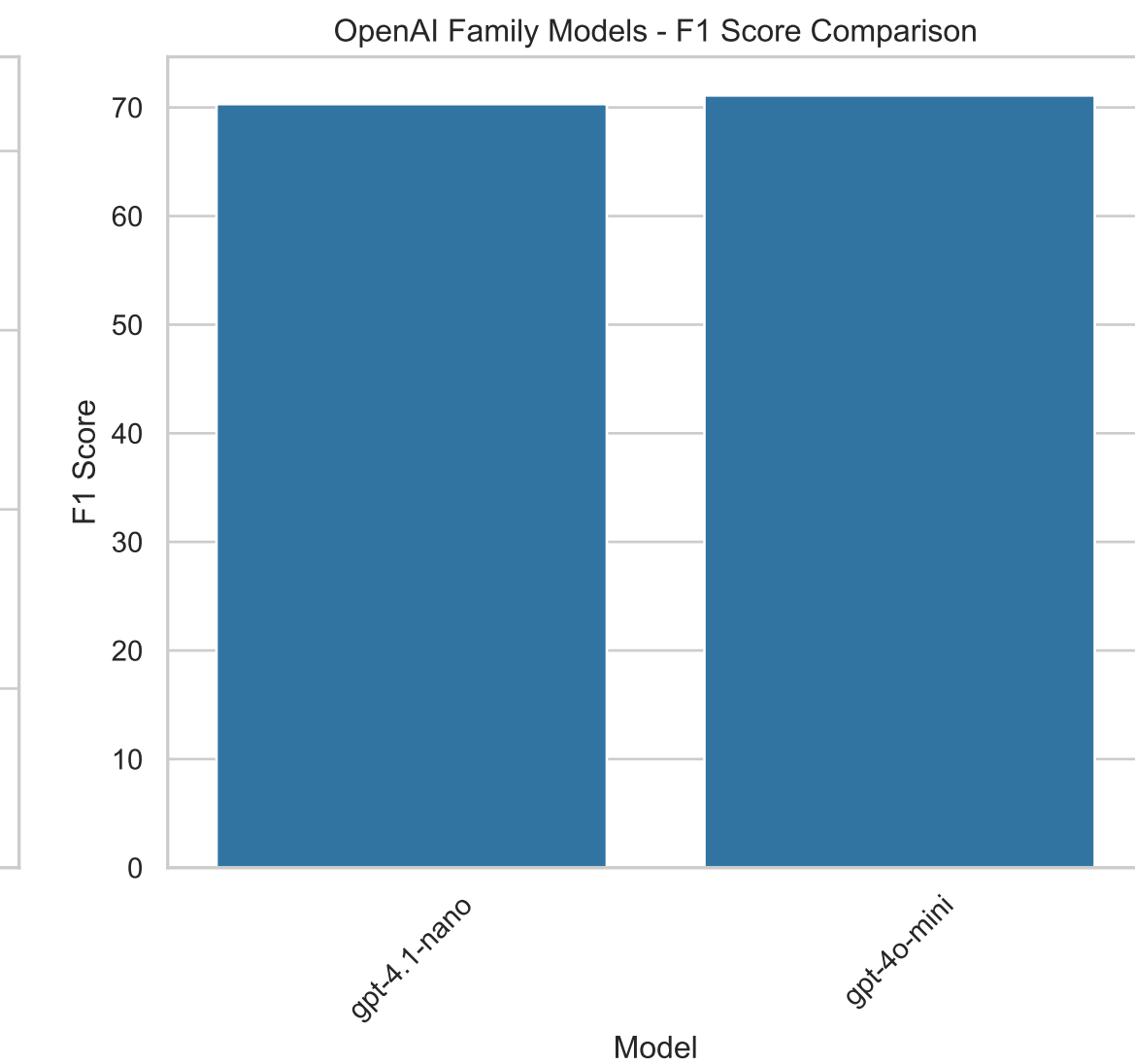
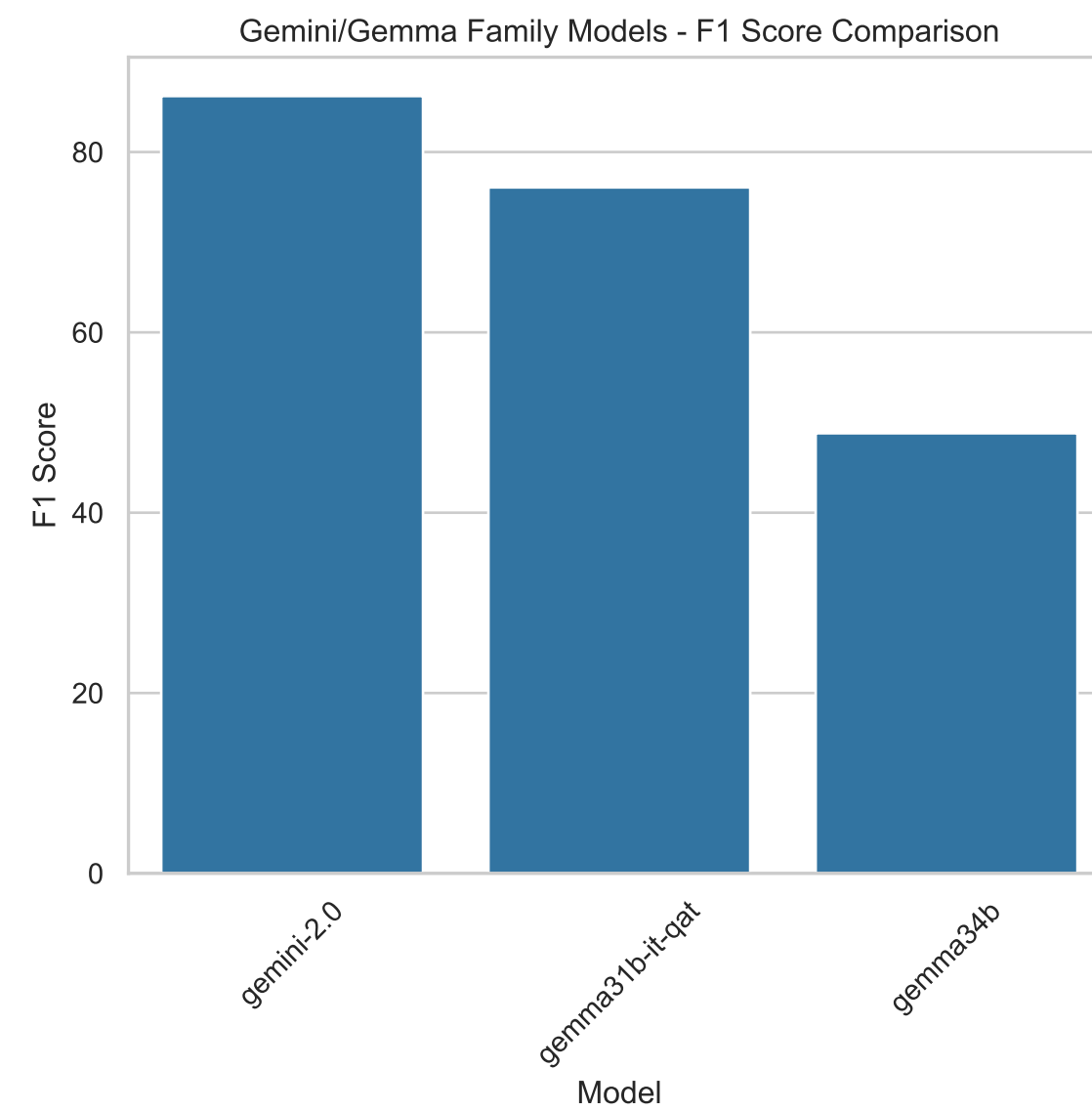
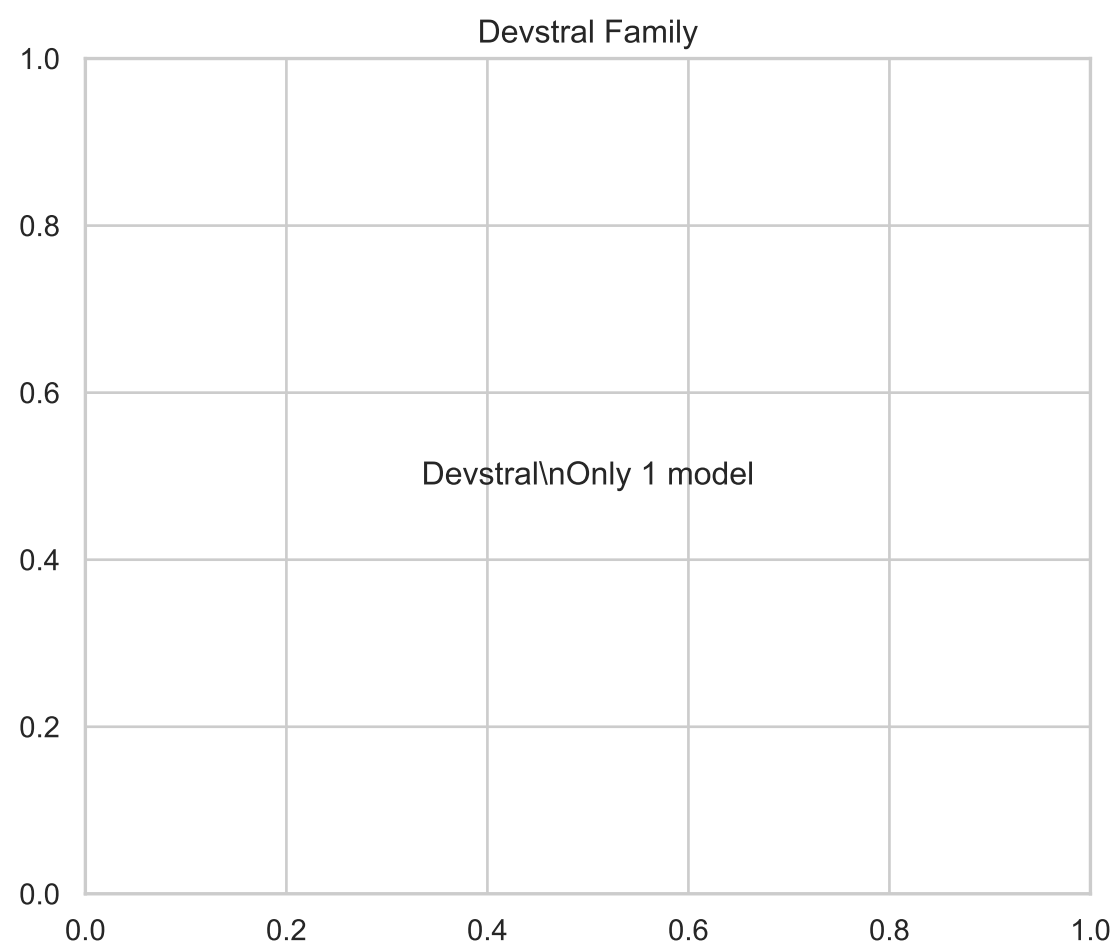


F1 Scores by Hospital and LLM Model

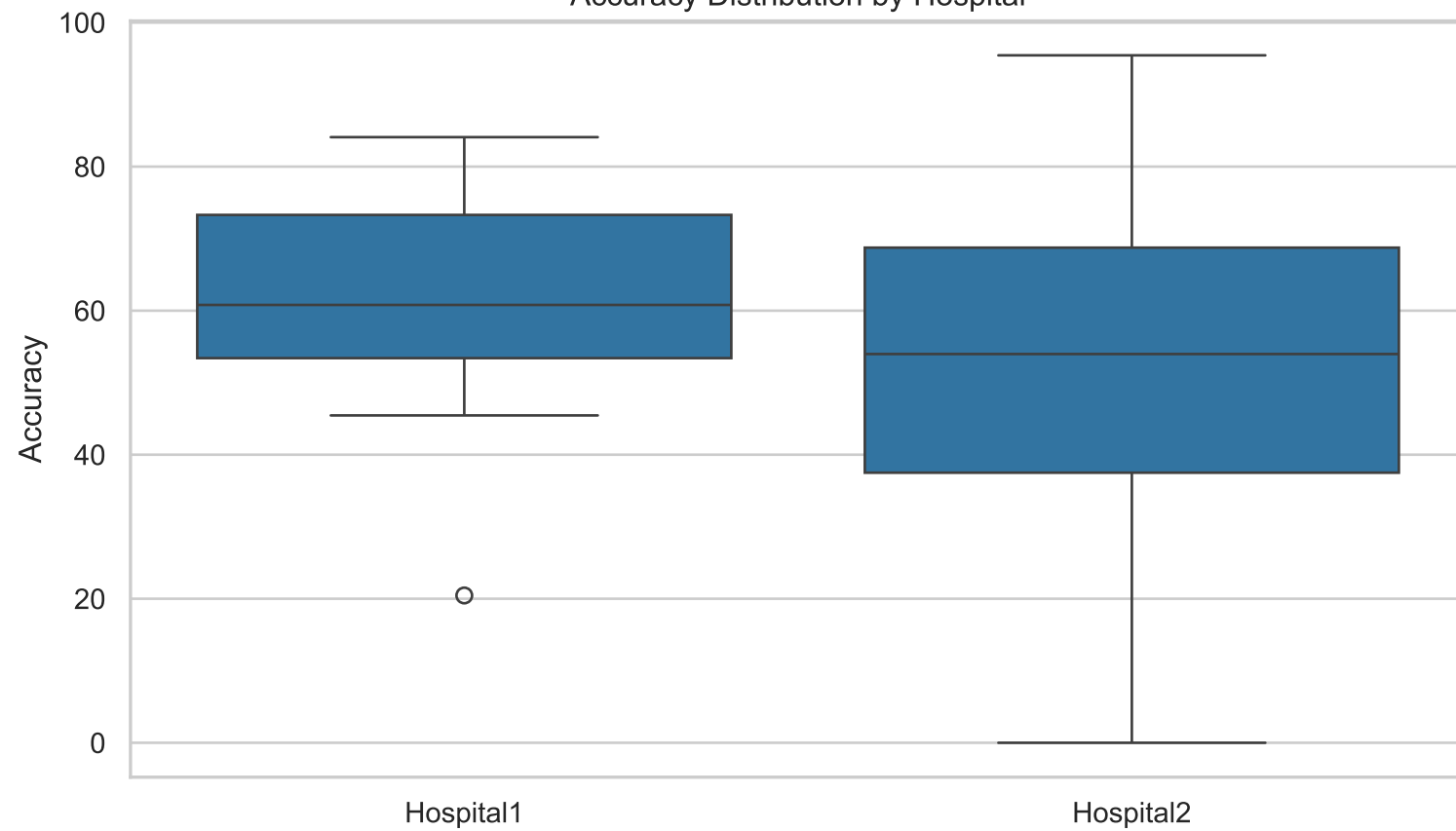


Average F1 Score by Model Family

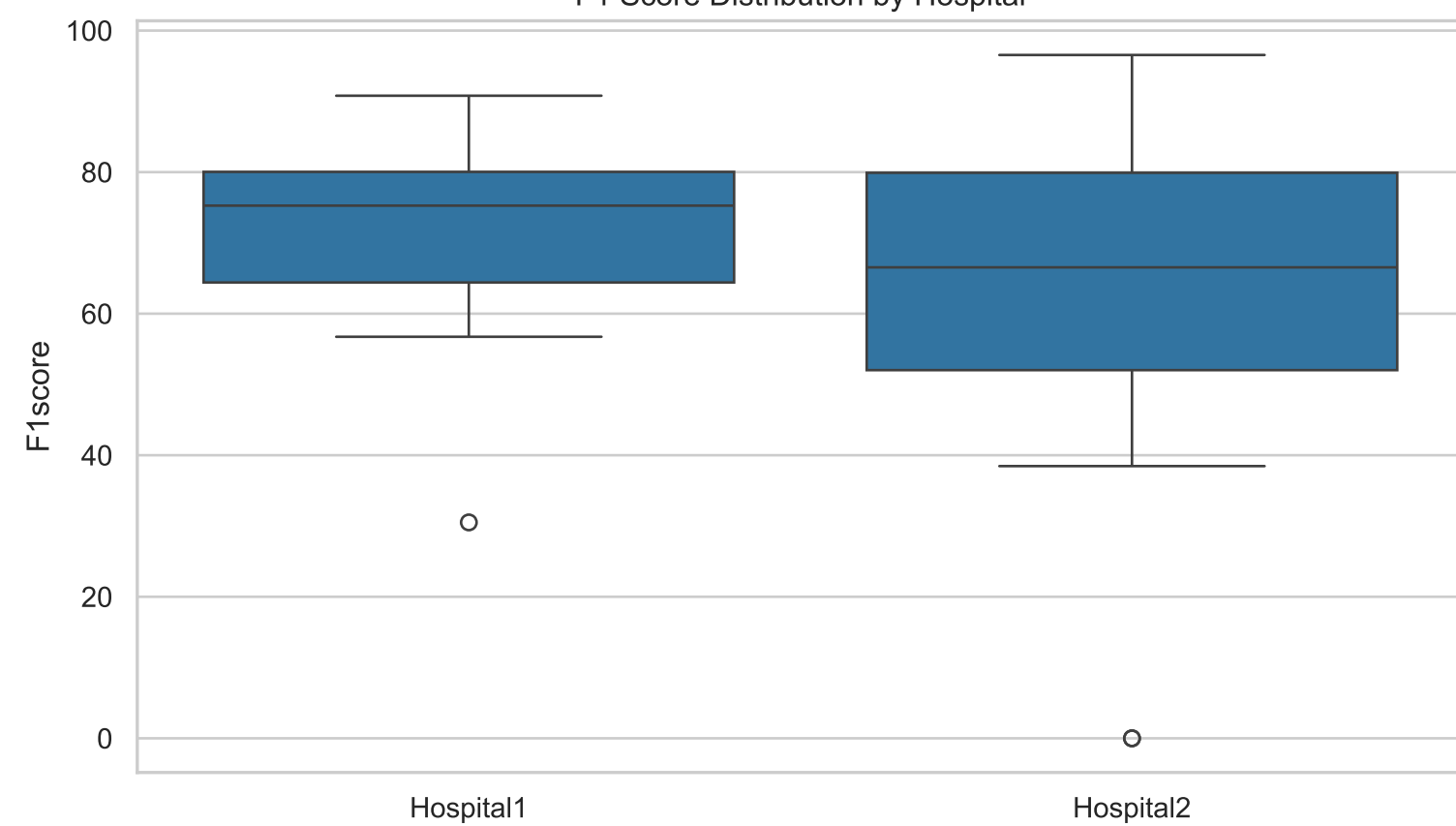




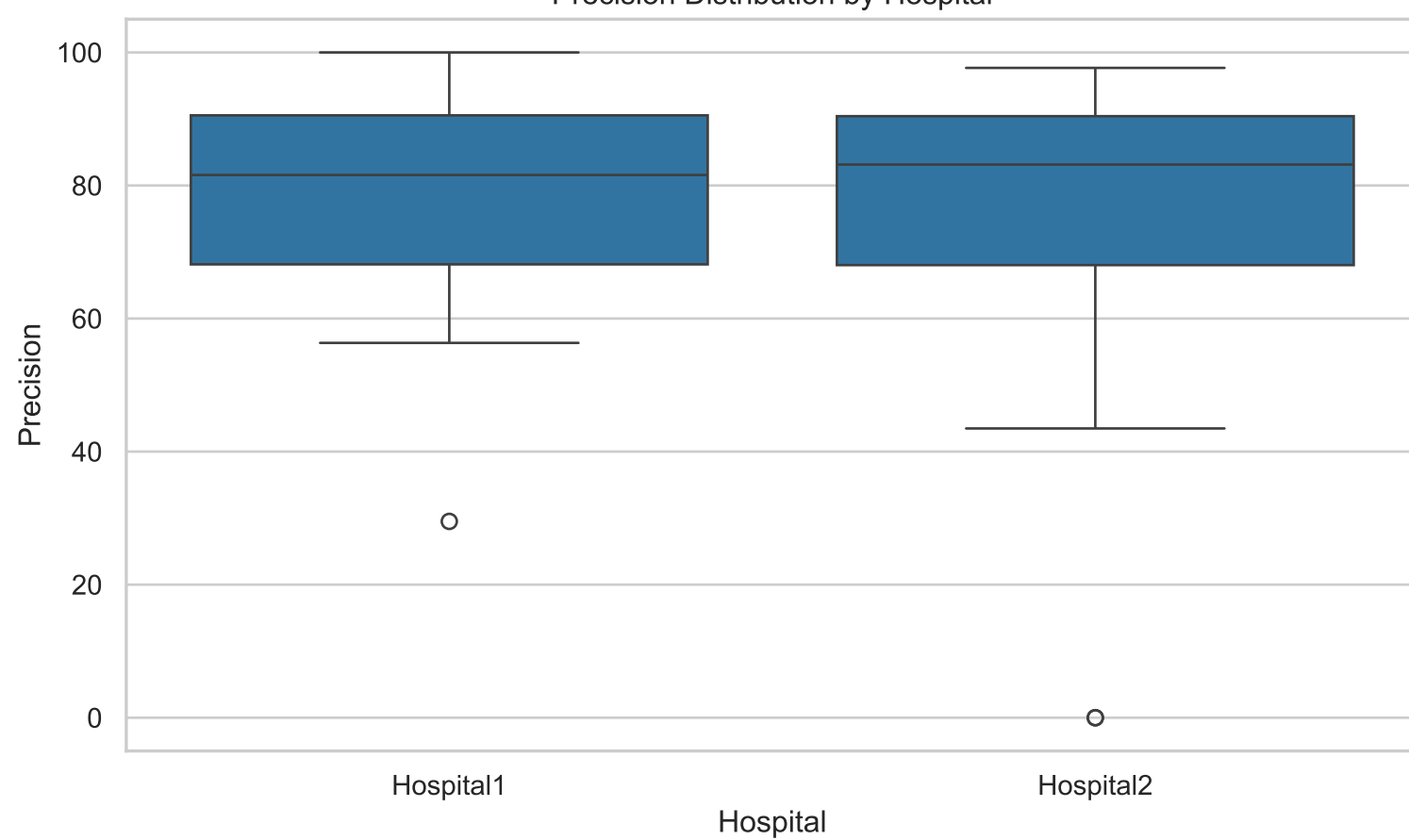
Accuracy Distribution by Hospital



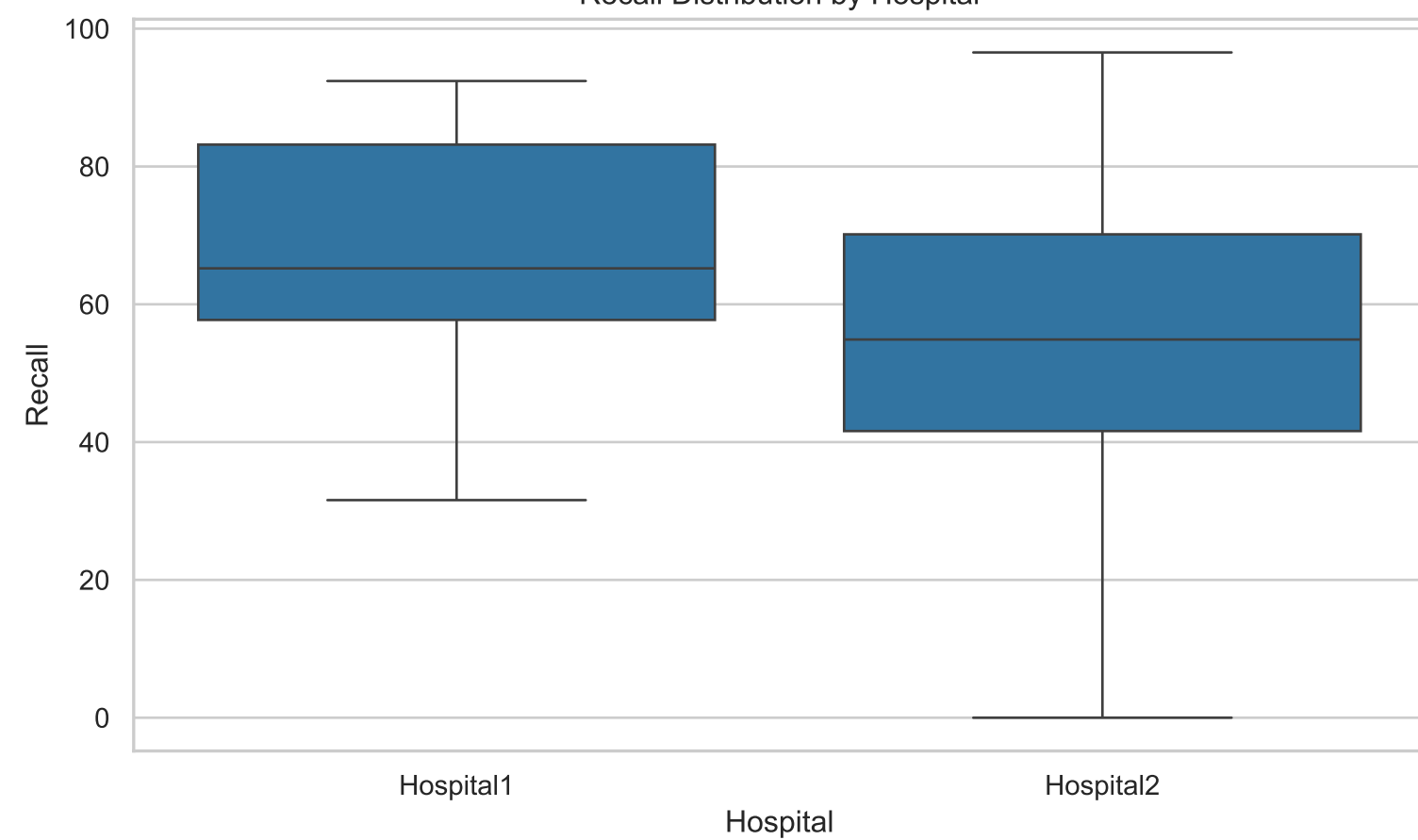
F1 Score Distribution by Hospital



Precision Distribution by Hospital

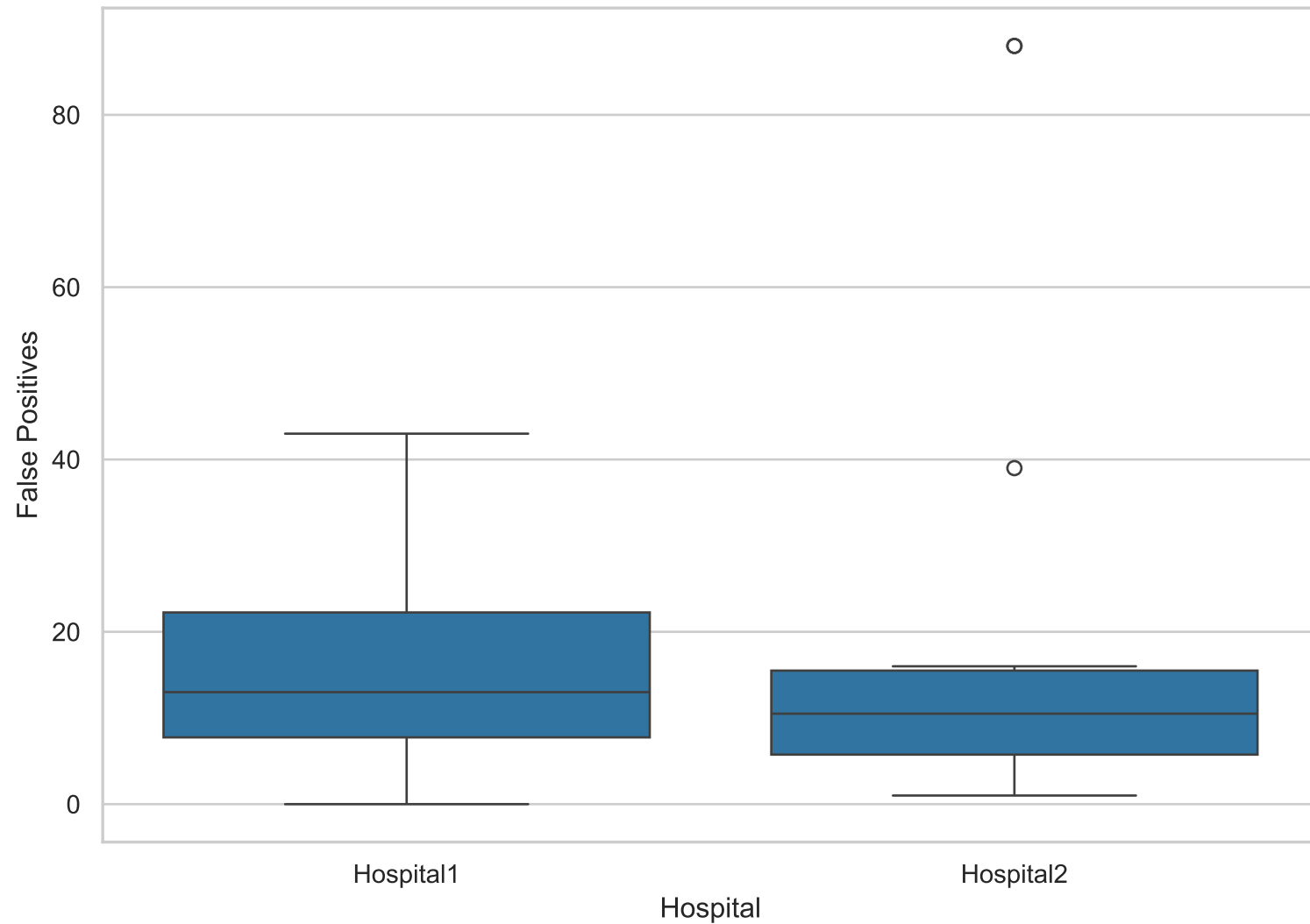


Recall Distribution by Hospital

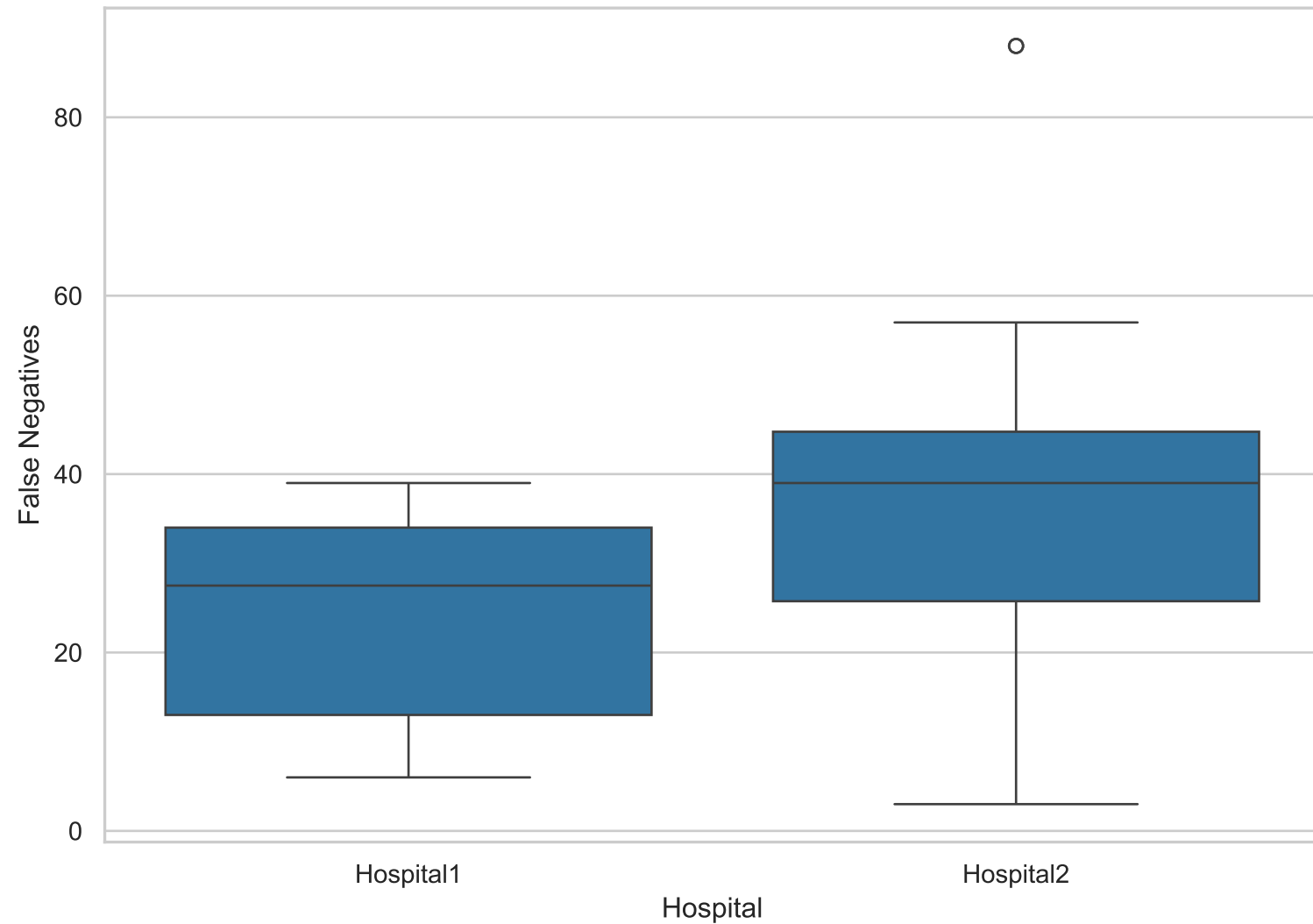




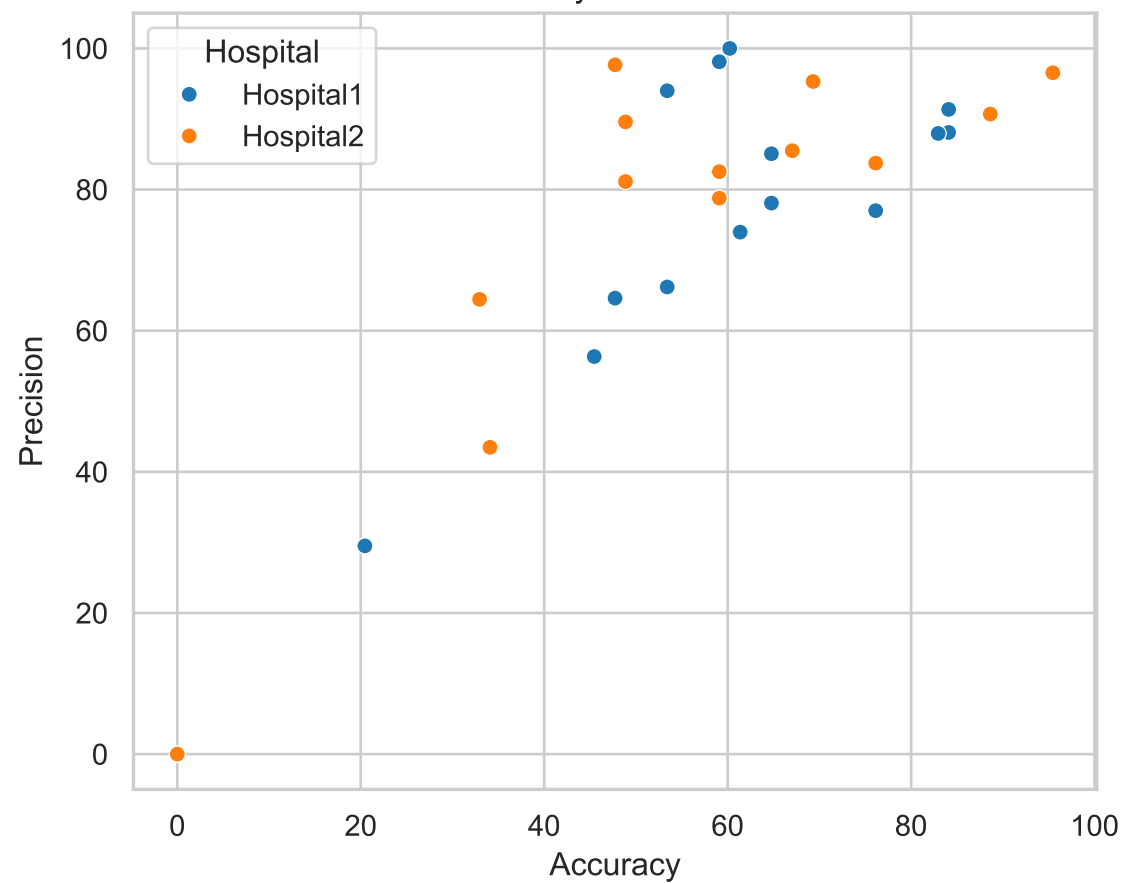
False Positives Distribution by Hospital



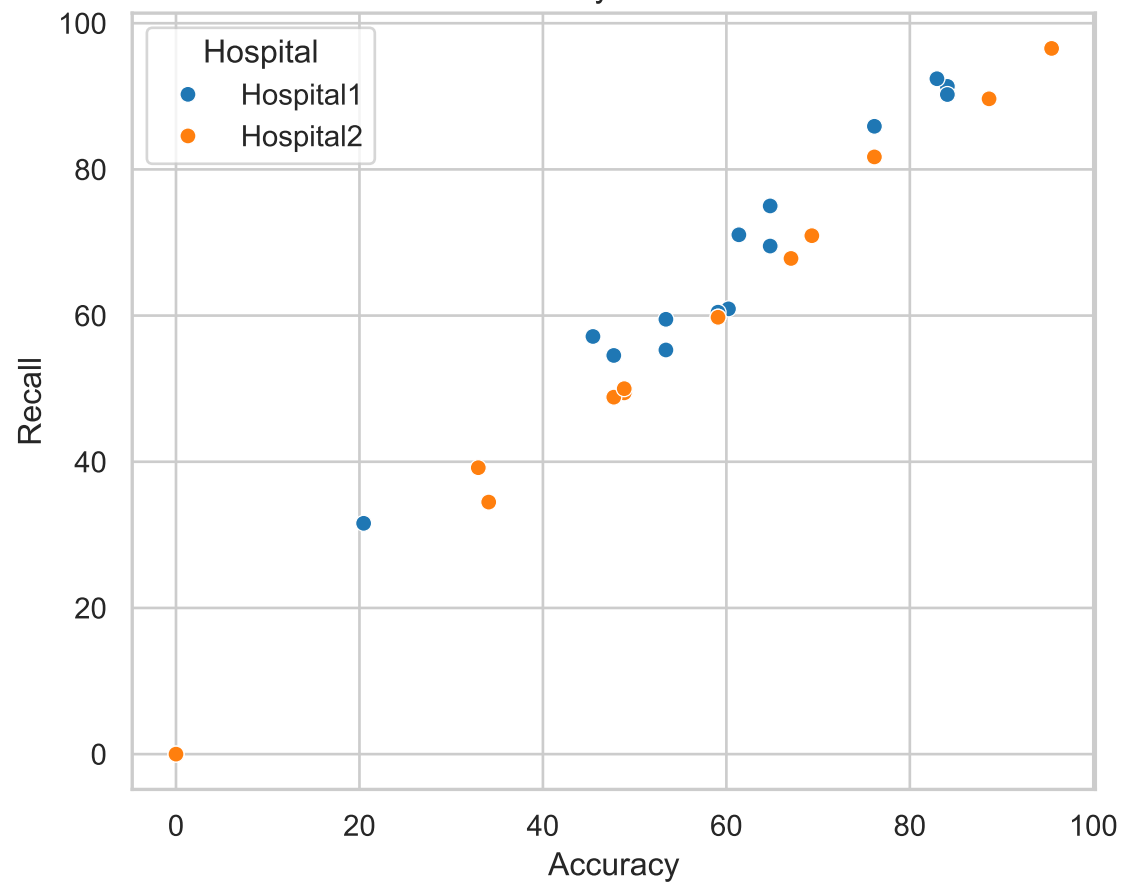
False Negatives Distribution by Hospital



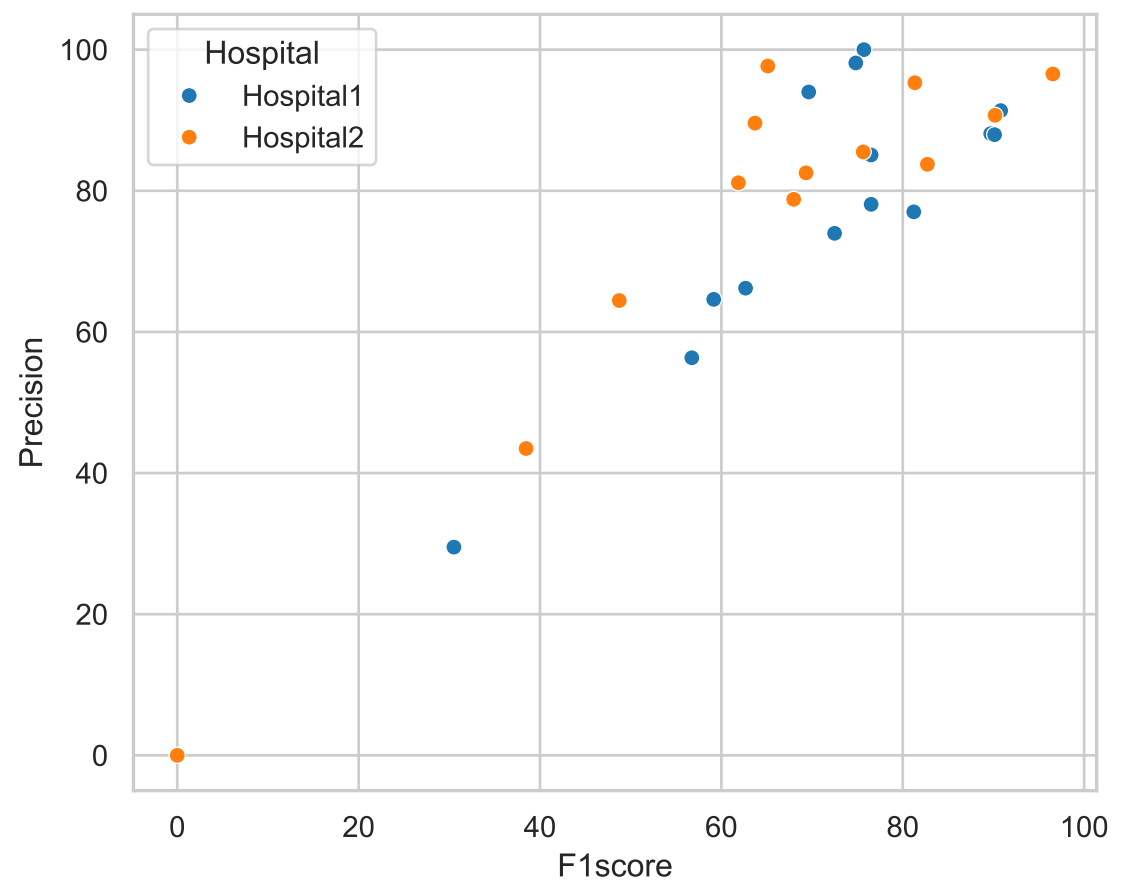
Accuracy vs Precision



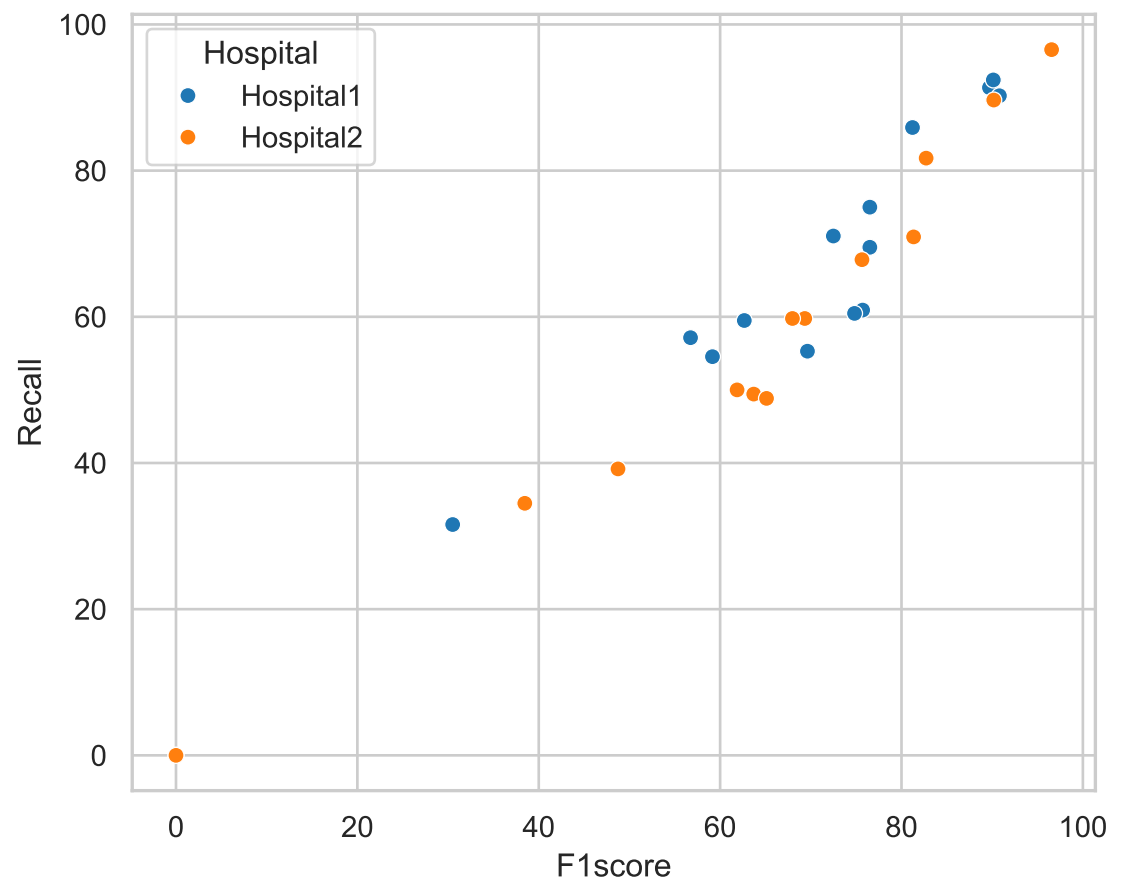
Accuracy vs Recall



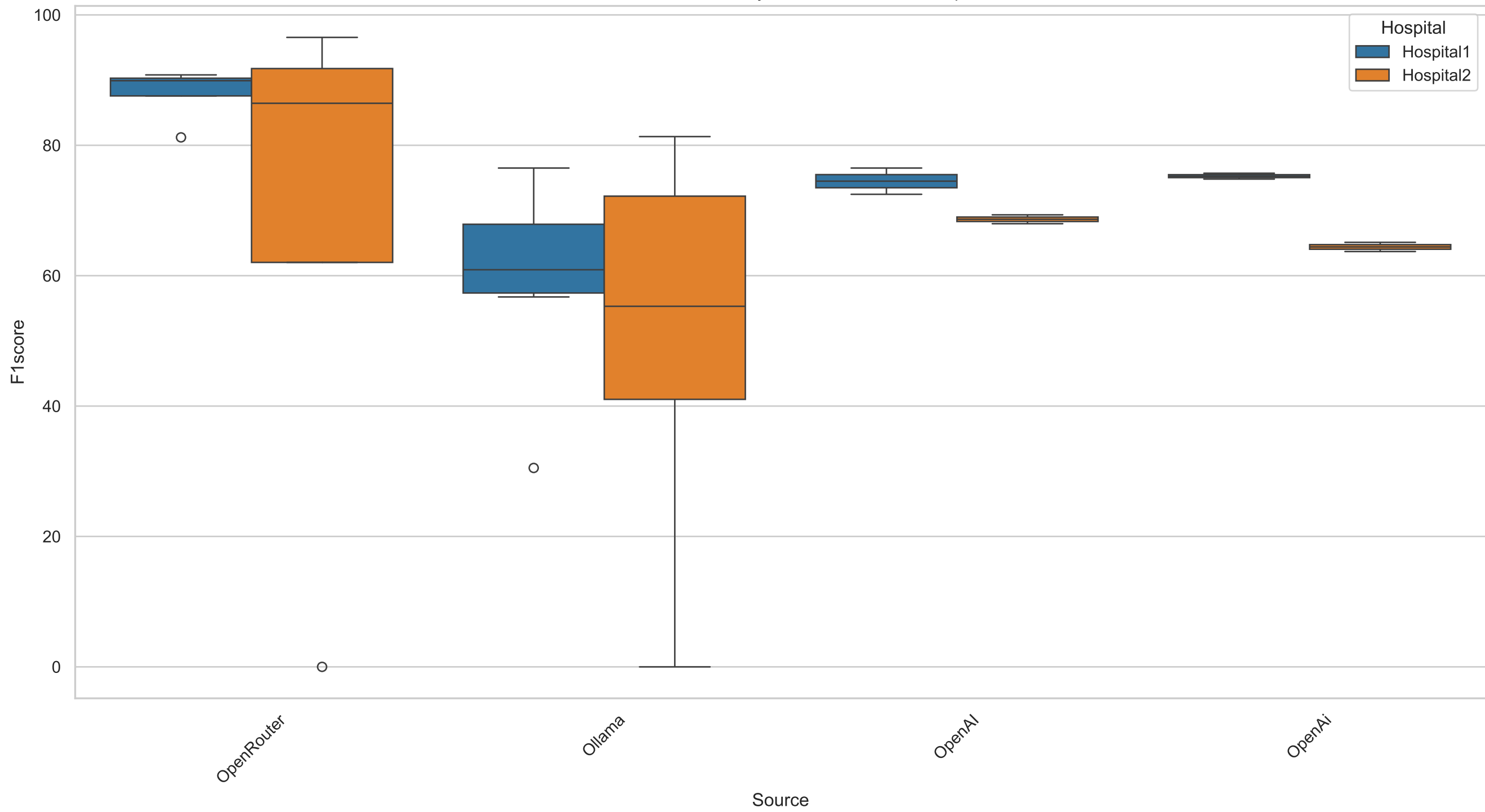
F1score vs Precision



F1score vs Recall



F1 Score Distribution by Model Source and Hospital



Average F1 Score Heatmap: Hospital vs Model Source

