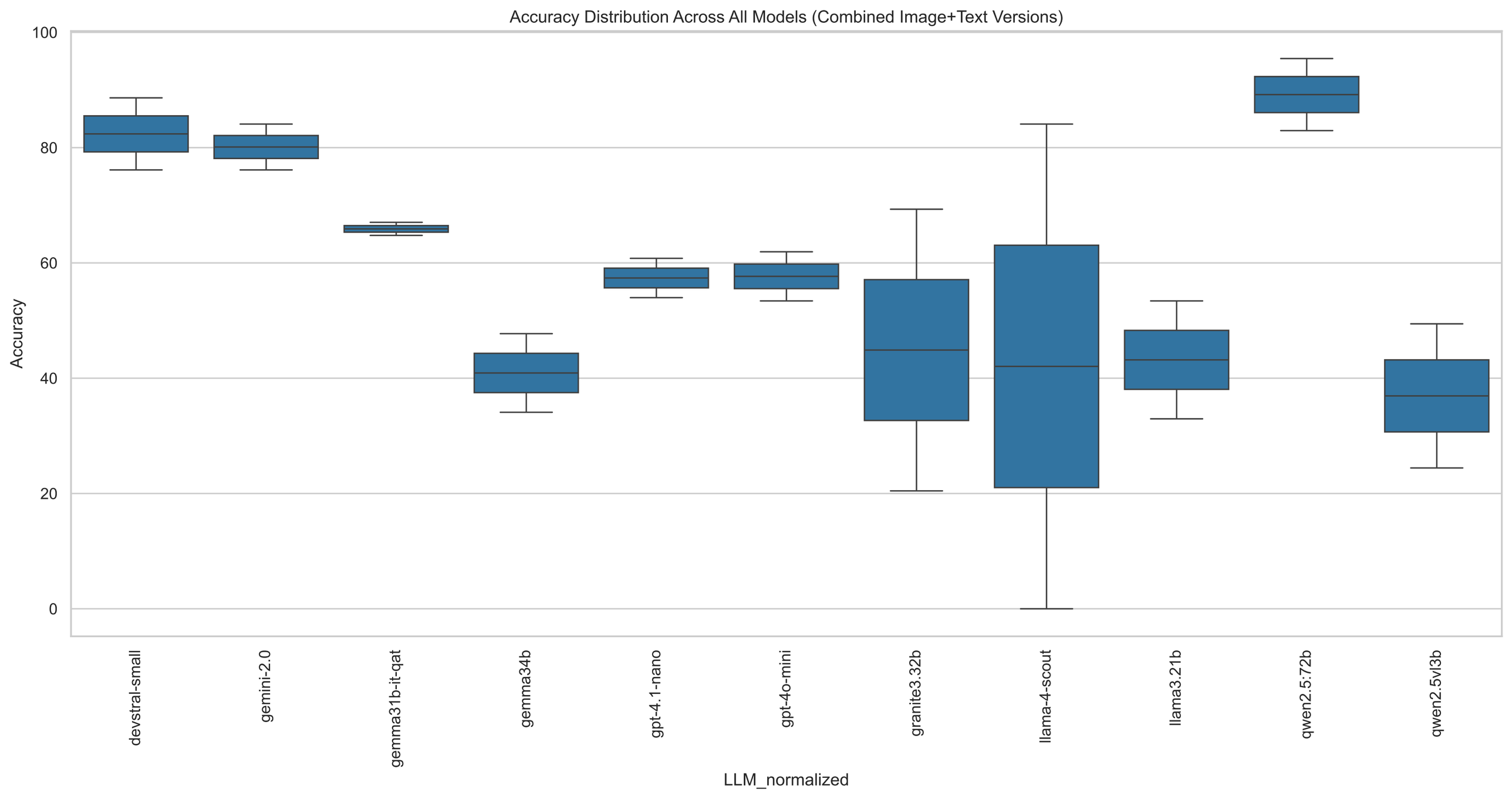
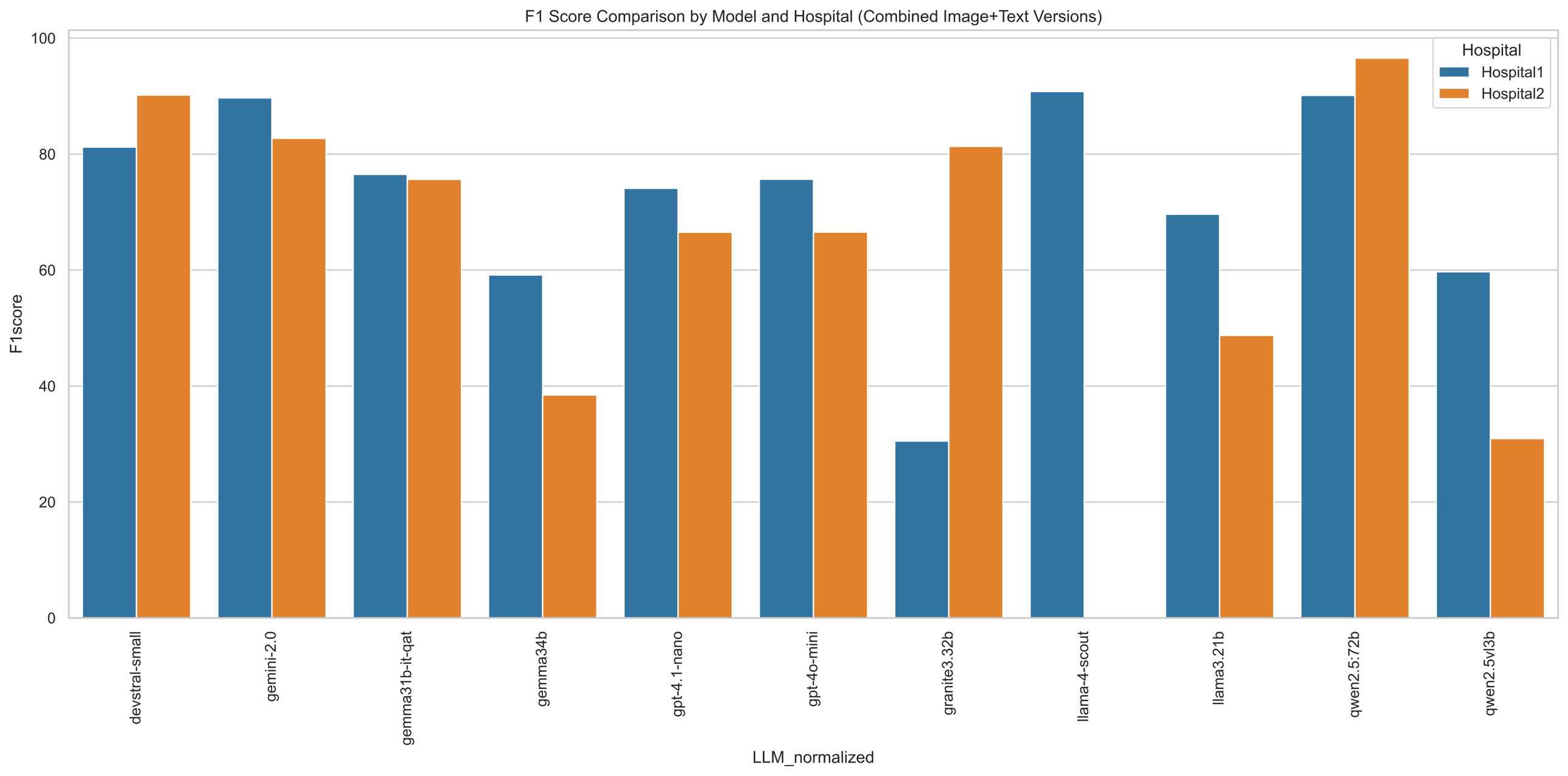
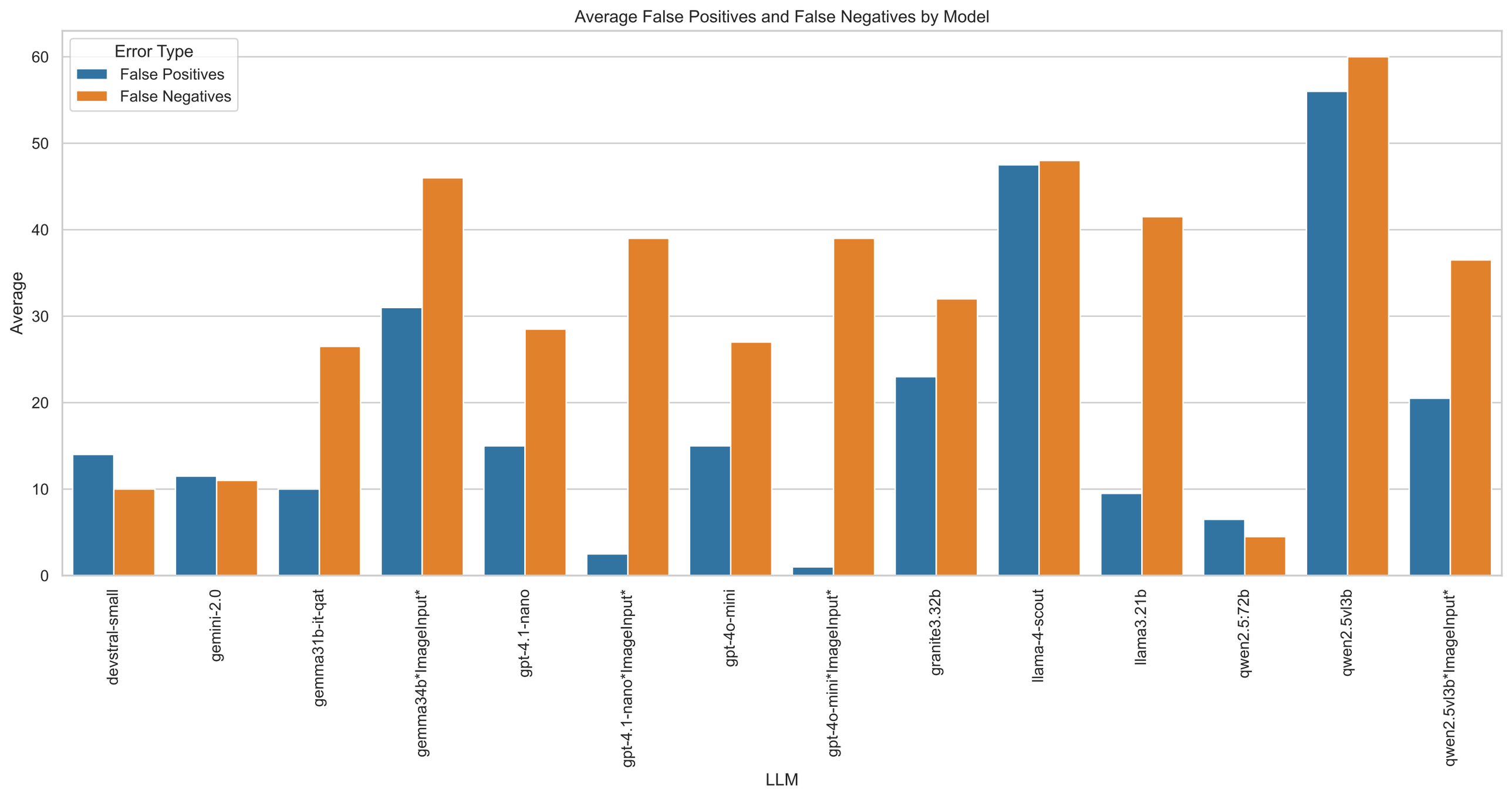
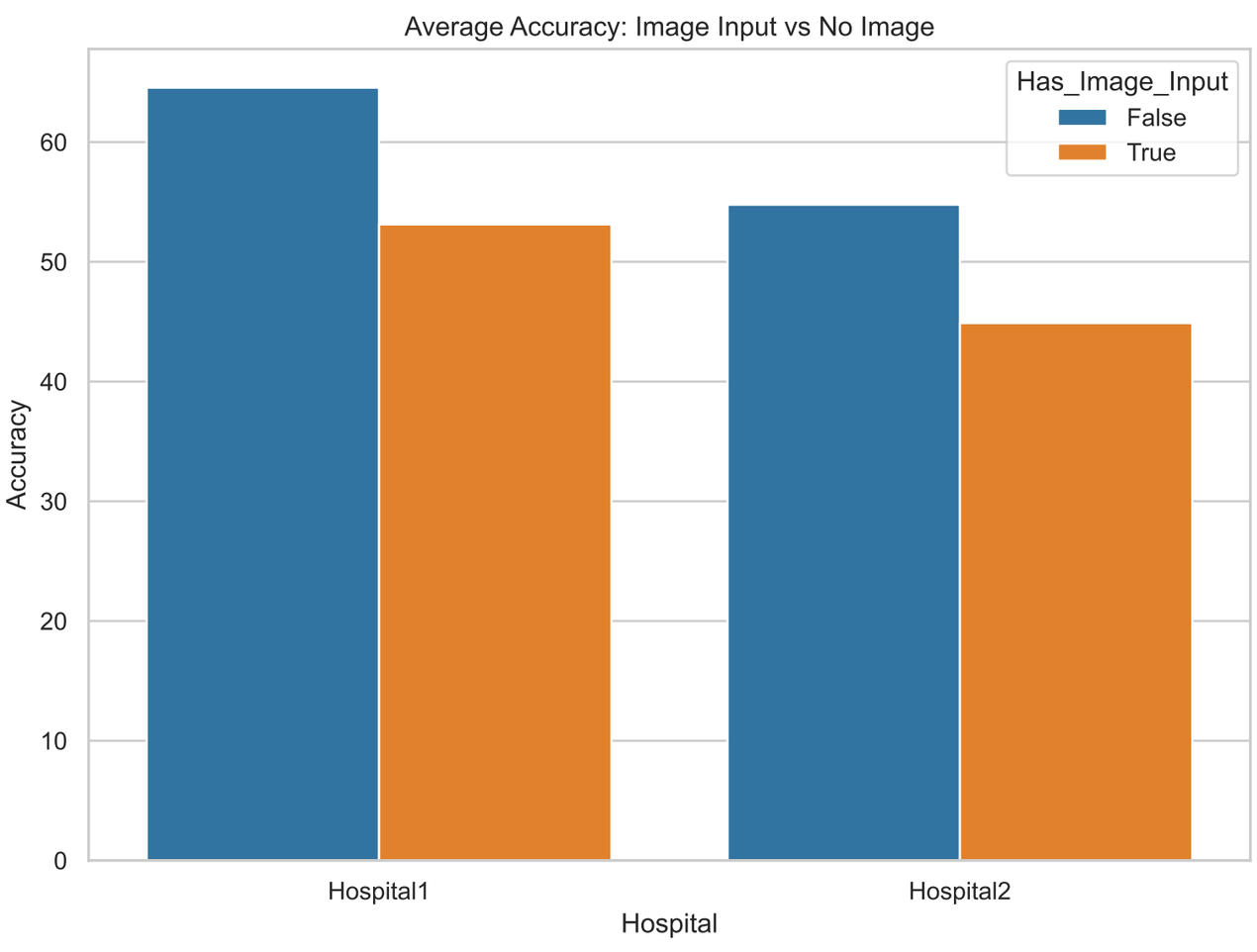
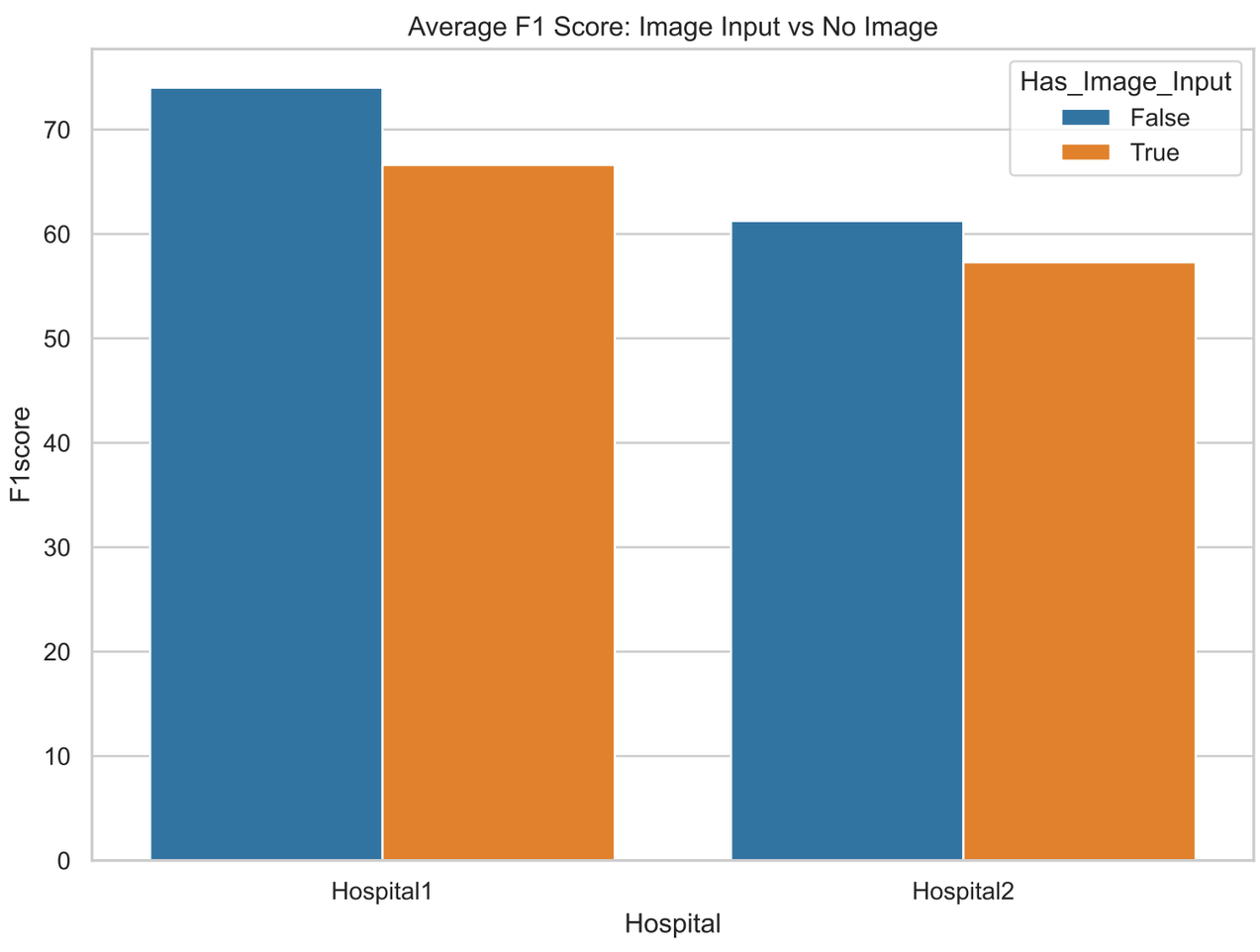
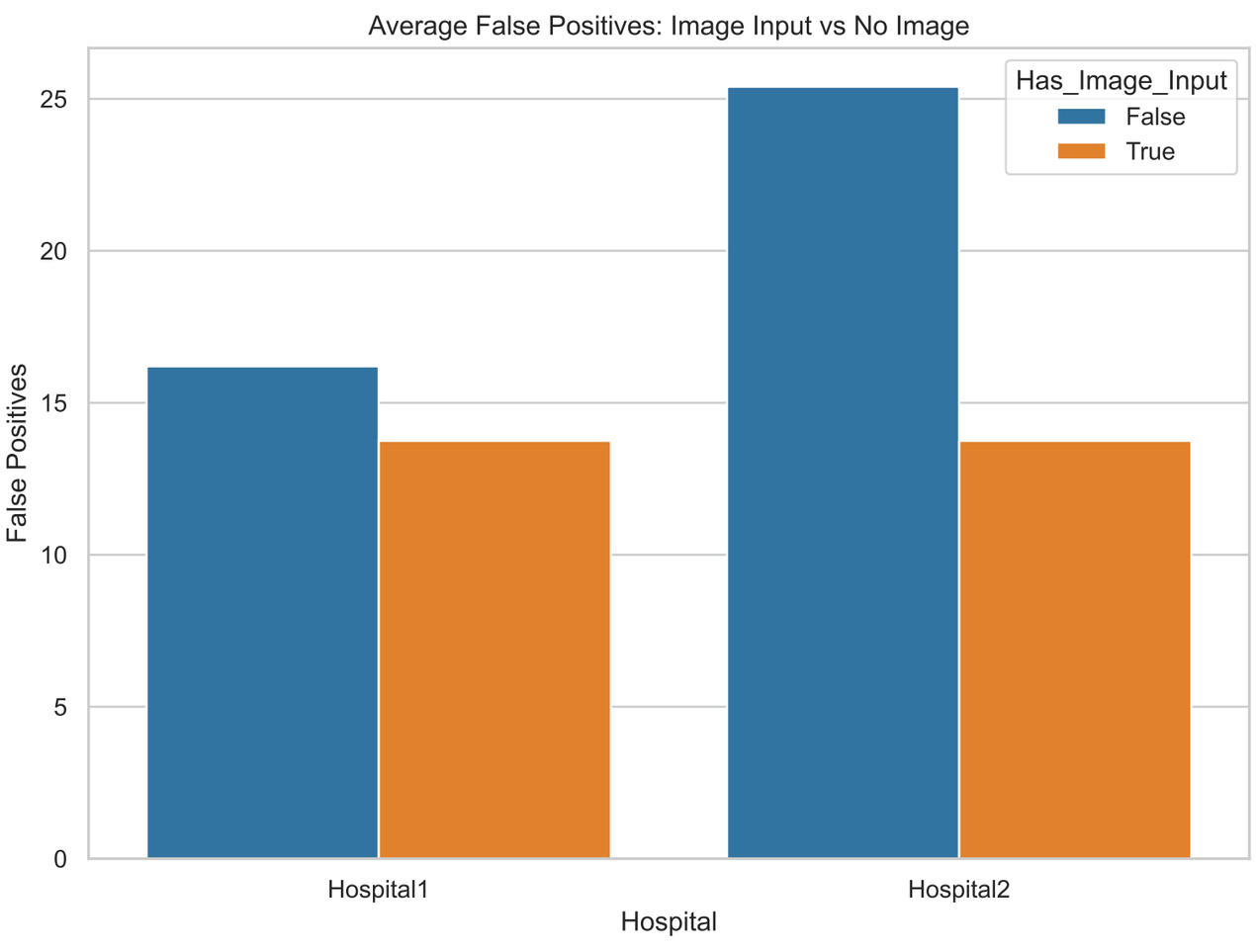

Accuracy Distribution Across All Models (Combined Image+Text Versions)
Accuracy
100
80
60
40
20
0
devstral-small
gemini-2.0
gemma31b-it-qat
gemma34b
gpt-4.1-nano
gpt-4o-mini
granite3.32b
llama-4-scout
llama3.21b
qwen2.5:72b
qwen2.5vl3b
LLM_normalized

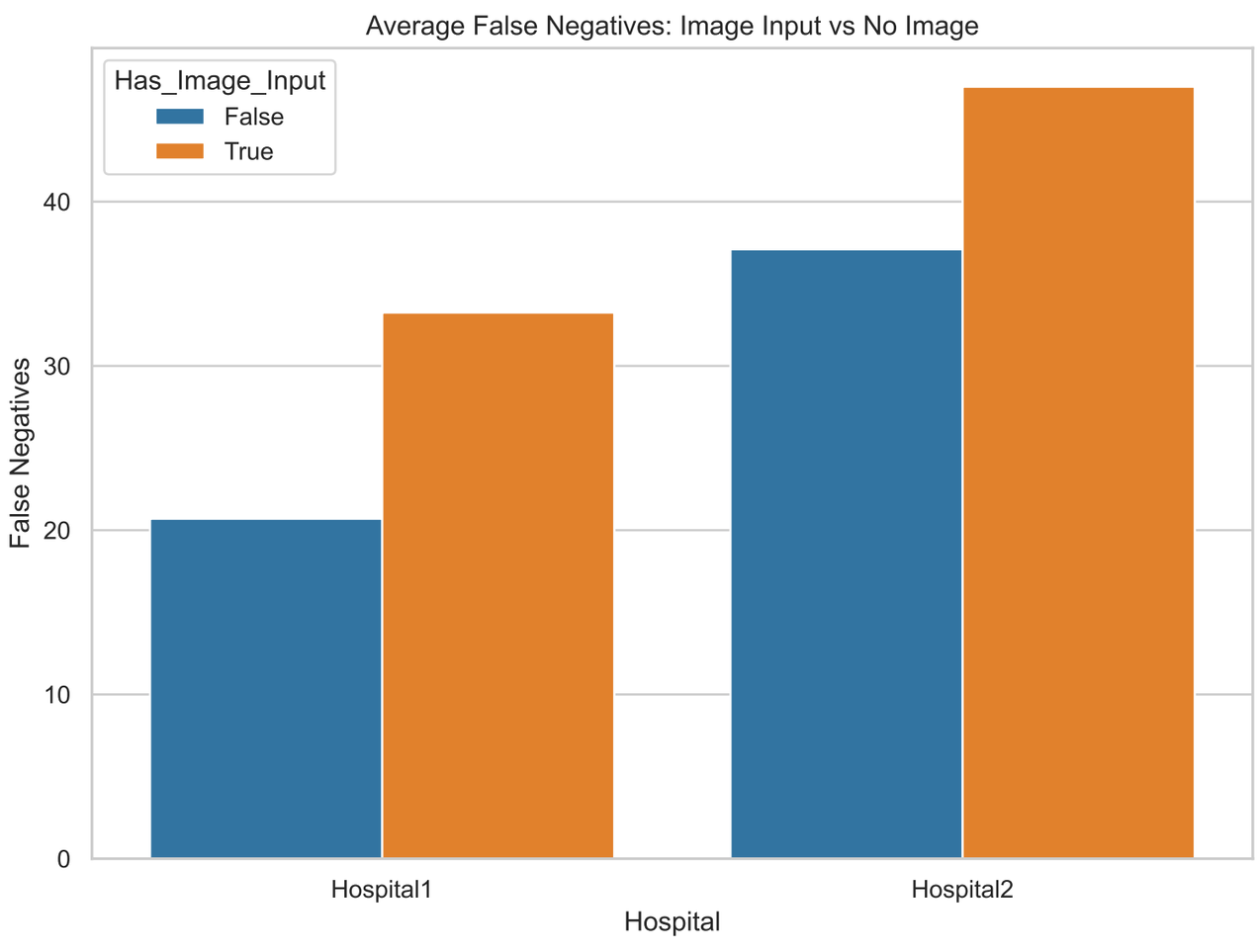

F1 Score Comparison by Model and Hospital (Combined Image+Text Versions)
Hospital
Hospital1
Hospital2
F1score
100
80
60
40
20
0
devstral-small
gemini-2.0
gemma31b-it-qat
gemma34b
gpt-4.1-nano
gpt-4o-mini
granite3.32b
llama-4-scout
llama3.21b
qwen2.5:72b
qwen2.5vl3b
LLM_normalized

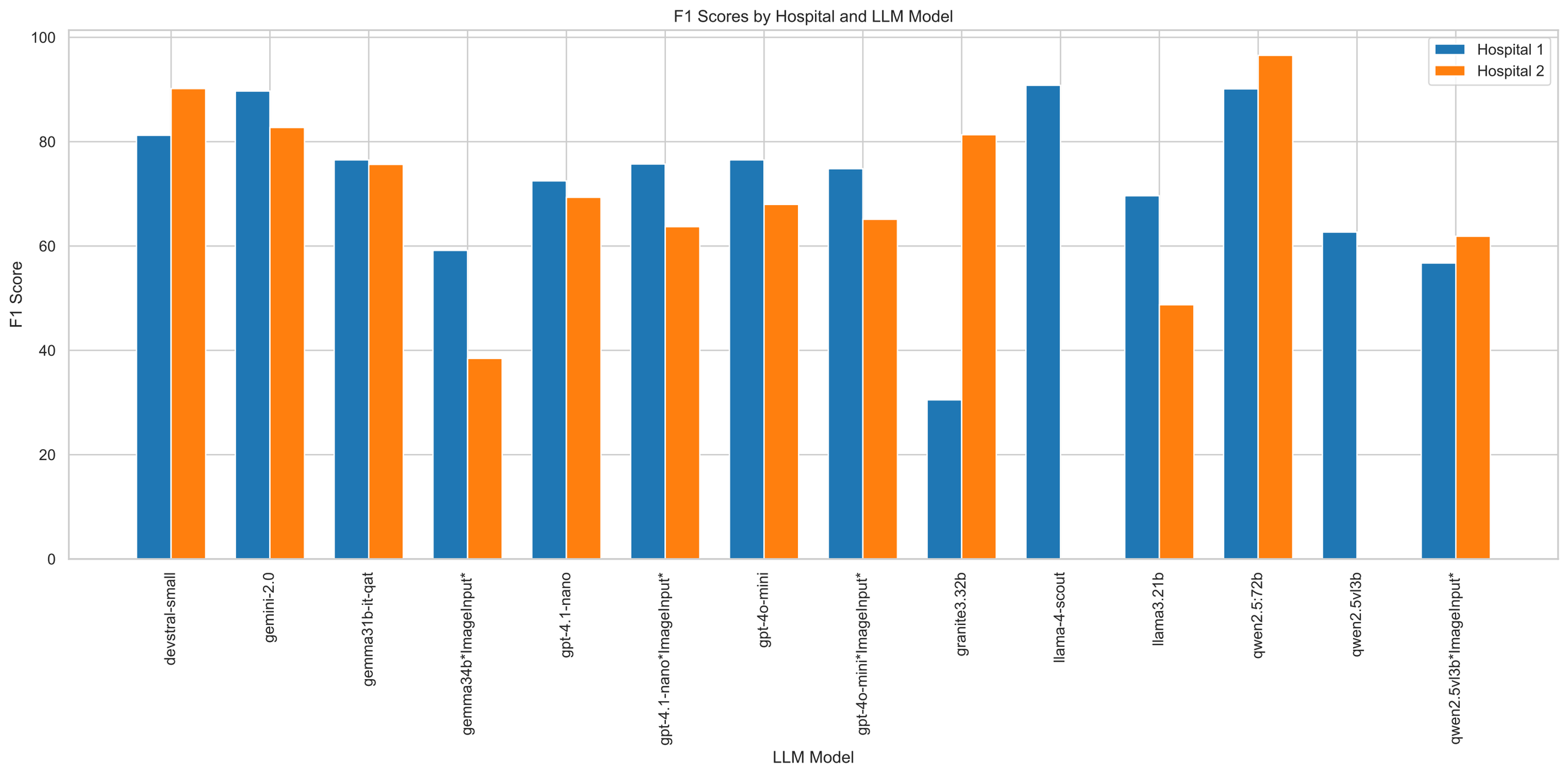

Average False Positives and False Negatives by Model
Error Type
False Positives
False Negatives
Average
0
10
20
30
40
50
60
devstral-small
gemini-2.0
gemma31b-it-qat
gemma34b*ImageInput*
gpt-4.1-nano
gpt-4.1-nano*ImageInput*
gpt-4o-mini
gpt-4o-mini*ImageInput*
granite3.32b
llama-4-scout
llama3.21b
qwen2.5:72b
qwen2.5vl3b
qwen2.5vl3b*ImageInput*
LLM

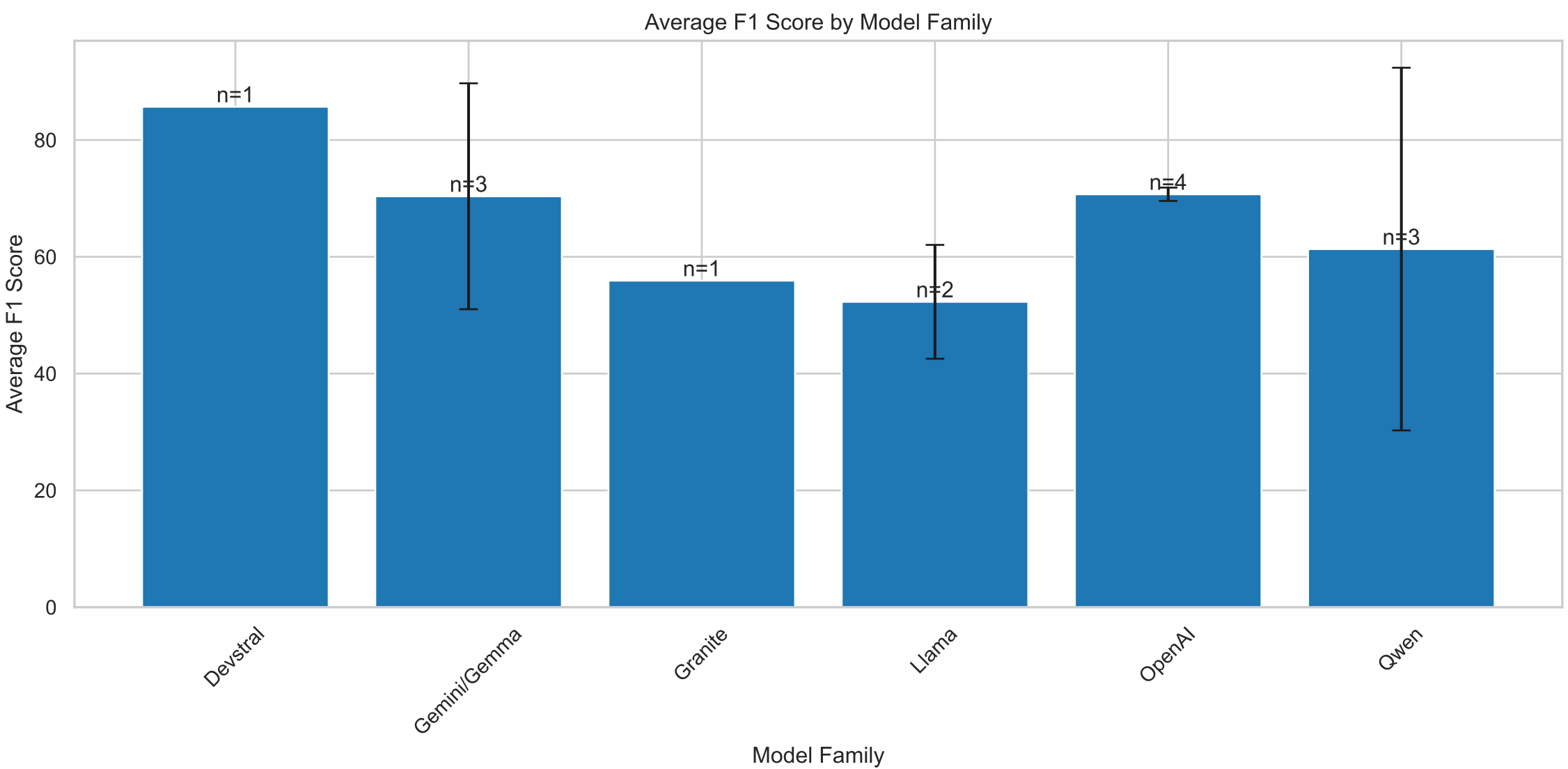

Average Accuracy: Image Input vs No Image
Has_Image_Input
False
True
Accuracy
0
10
20
30
40
50
60
Hospital1
Hospital2
Hospital

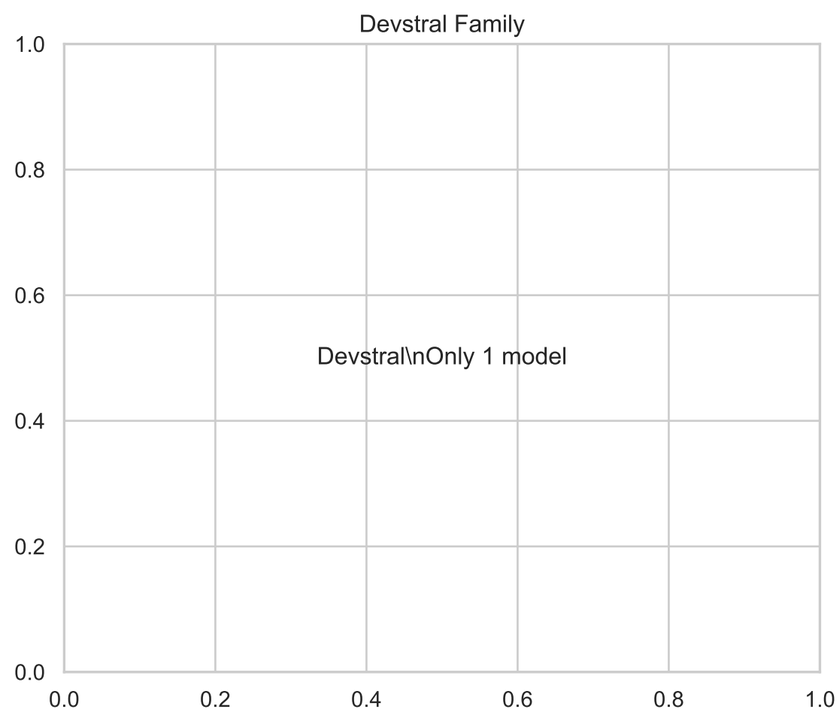

Average F1 Score: Image Input vs No Image
Has_Image_Input
False
True
F1score
0
10
20
30
40
50
60
70
Hospital1
Hospital2
Hospital

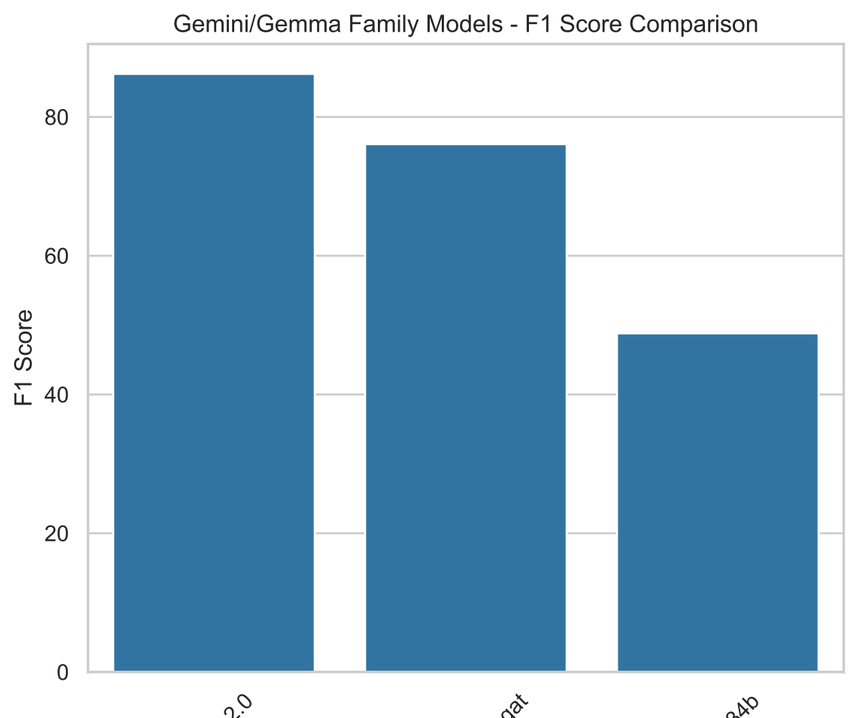

Average False Positives: Image Input vs No Image
Has_Image_Input
False
True
False Positives
0
5
10
15
20
25
Hospital1
Hospital2
Hospital


Average False Negatives: Image Input vs No Image
Has_Image_Input
False
True
False Negatives
0
10
20
30
40
Hospital1
Hospital2
Hospital

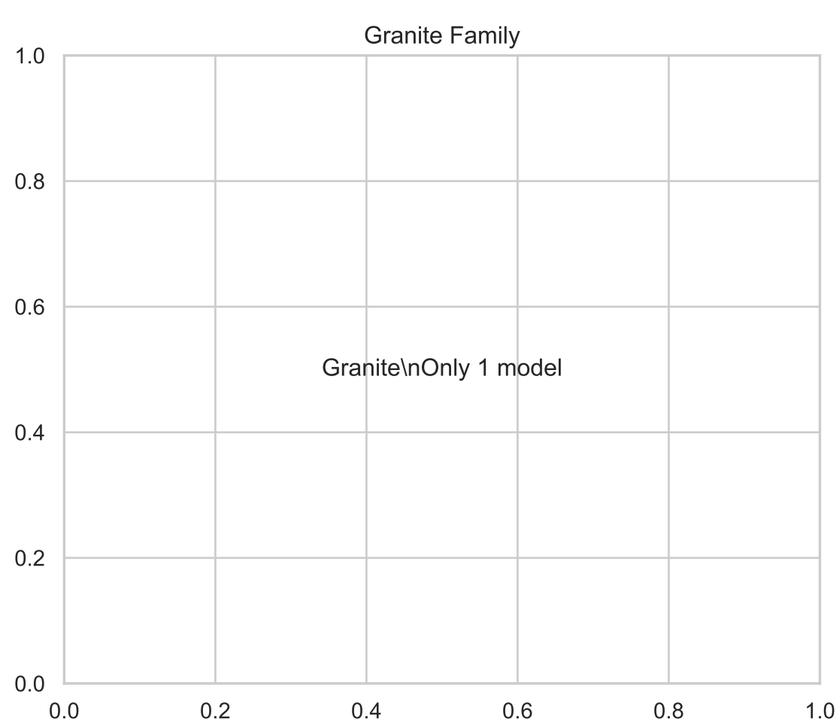

F1 Scores by Hospital and LLM Model
Hospital 1
Hospital 2
F1 Score
0
20
40
60
80
100
devstral-small
gemini-2.0
gemma31b-it-qat
gemma34b*ImageInput*
gpt-4.1-nano
gpt-4.1-nano*ImageInput*
gpt-4o-mini
gpt-4o-mini*ImageInput*
granite3.32b
llama-4-scout
llama3.21b
qwen2.5:72b
qwen2.5vl3b
qwen2.5vl3b*ImageInput*
LLM Model

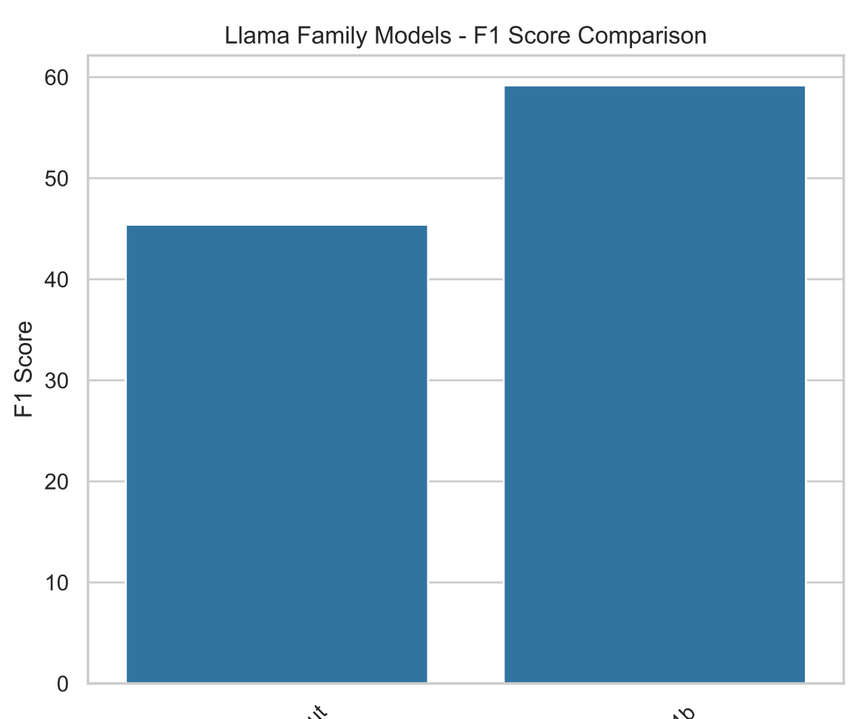

Average F1 Score by Model Family
Average F1 Score
0
20
40
60
80
n=1
n=3
n=1
n=2
n=4
n=3
Devstral
Gemini/Gemma
Granite
Llama
OpenAI
Qwen
Model Family

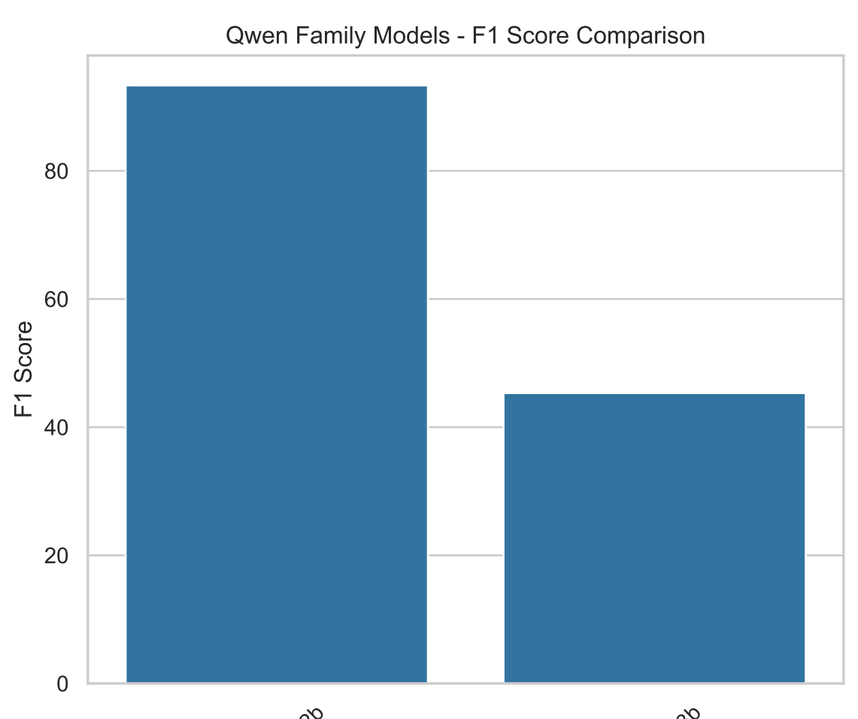

Devstral Family
Devstral\nOnly 1 model
0.0
0.2
0.4
0.6
0.8
1.0

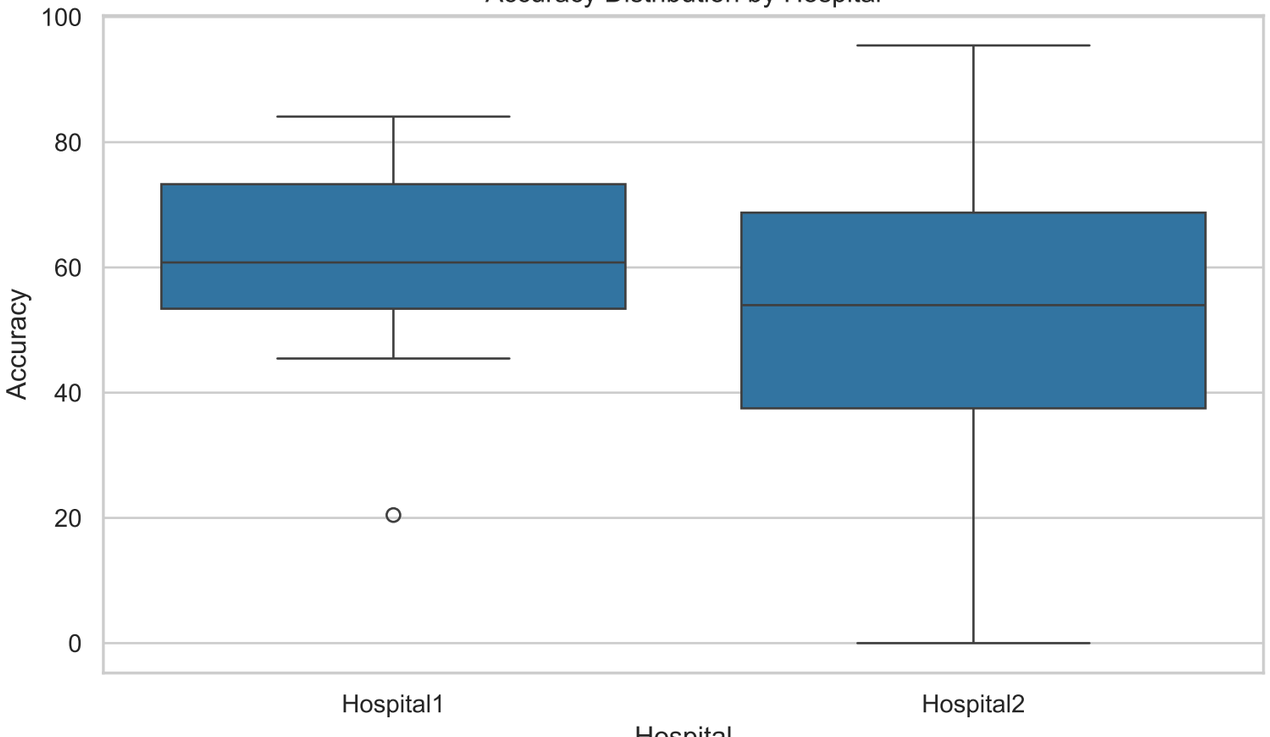

Gemini/Gemma Family Models - F1 Score Comparison
F1 Score
0
20
40
60
80
gemini-2.0
gemma31b-it-qat
gemma34b
Model

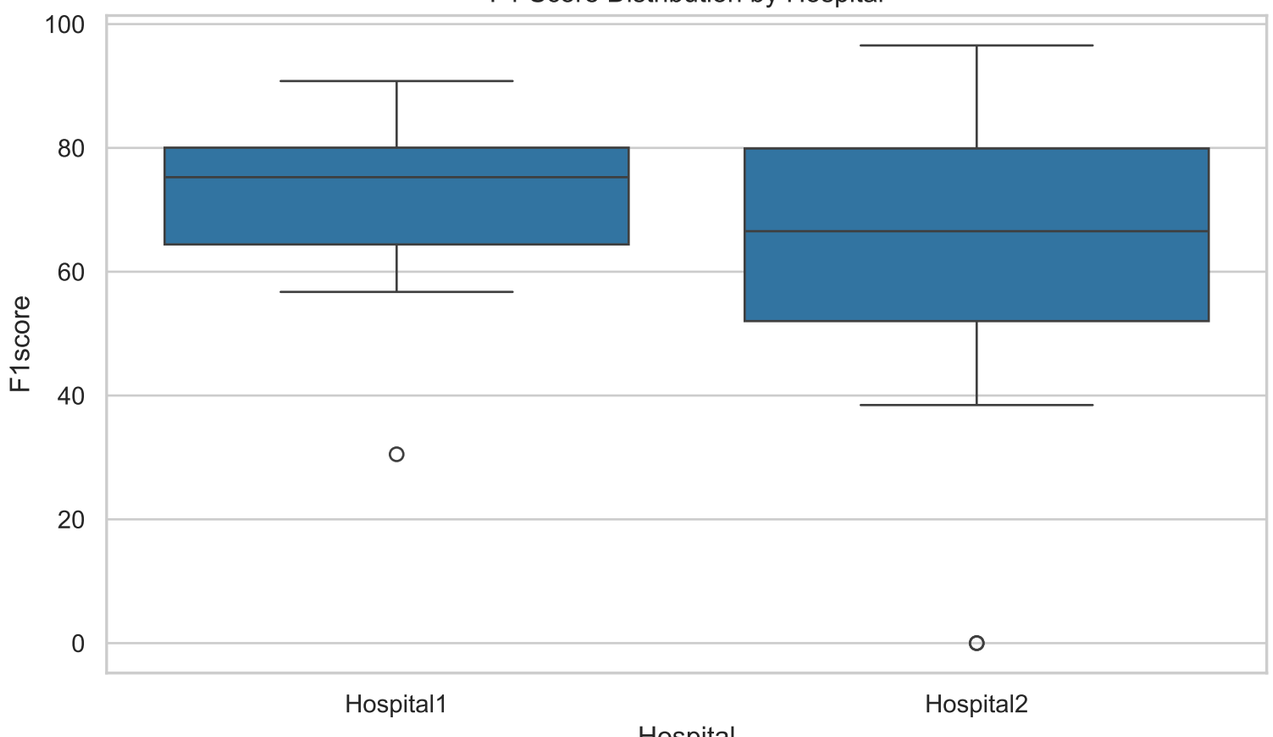

OpenAI Family Models - F1 Score Comparison
F1 Score
0
10
20
30
40
50
60
70
gpt-4.1-nano
gpt-4o-mini
Model

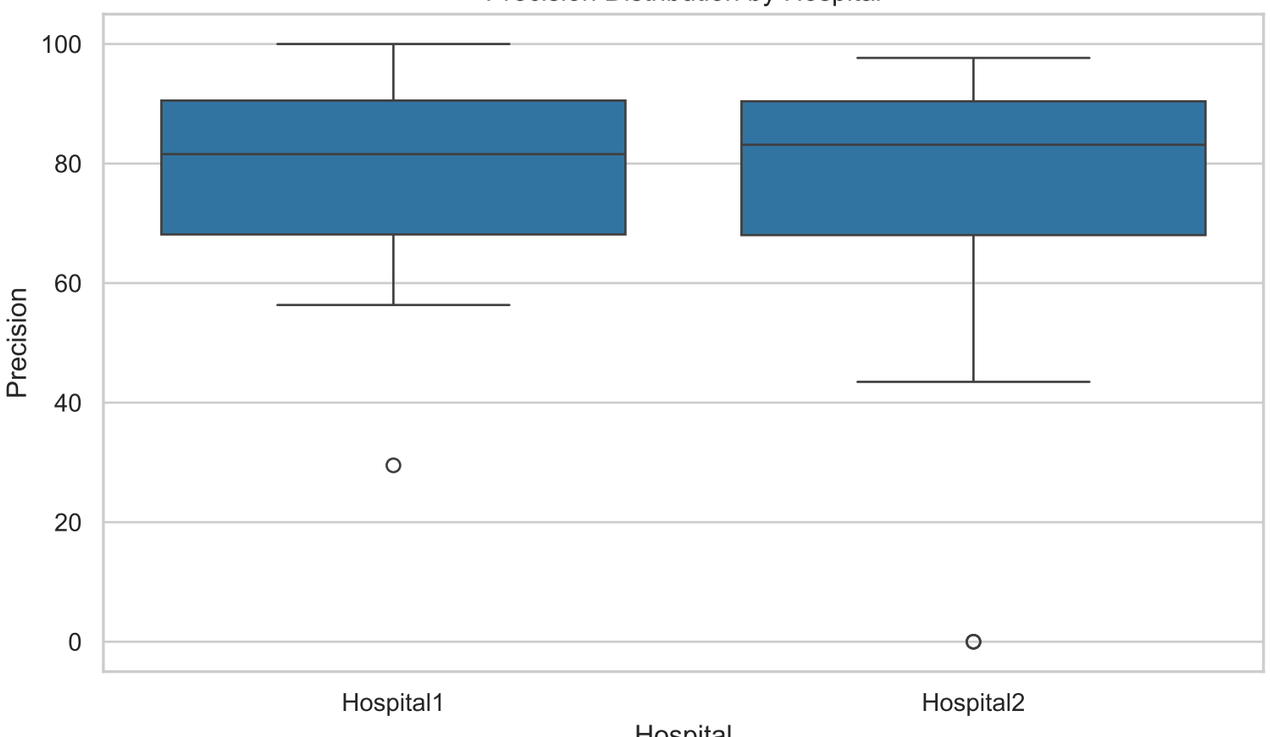

Granite Family
Granite\nOnly 1 model
0.0
0.2
0.4
0.6
0.8
1.0

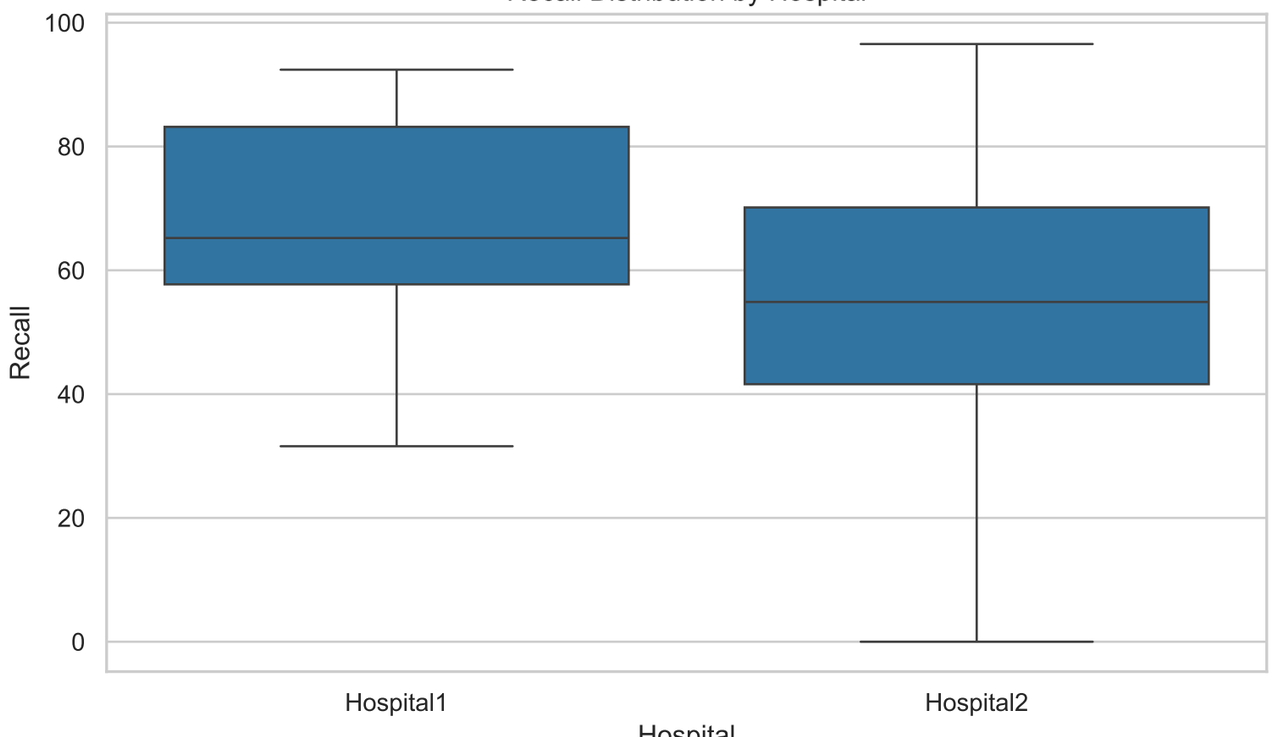

Llama Family Models - F1 Score Comparison
F1 Score
0
10
20
30
40
50
60
llama-4-scout
llama3.21b
Model

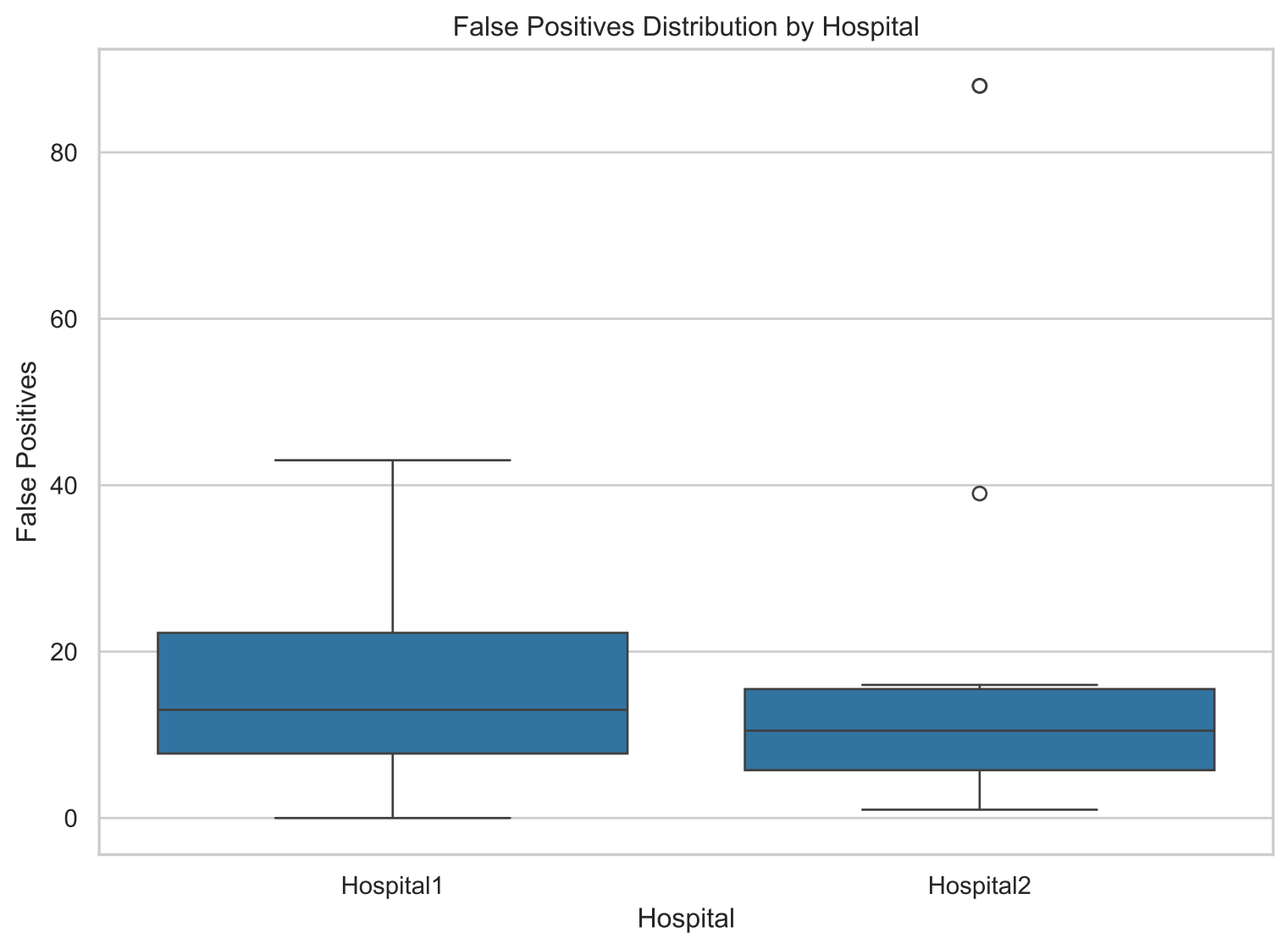

Qwen Family Models - F1 Score Comparison
F1 Score
0
20
40
60
80
qwen2.5:72b
qwen2.5vl3b
Model

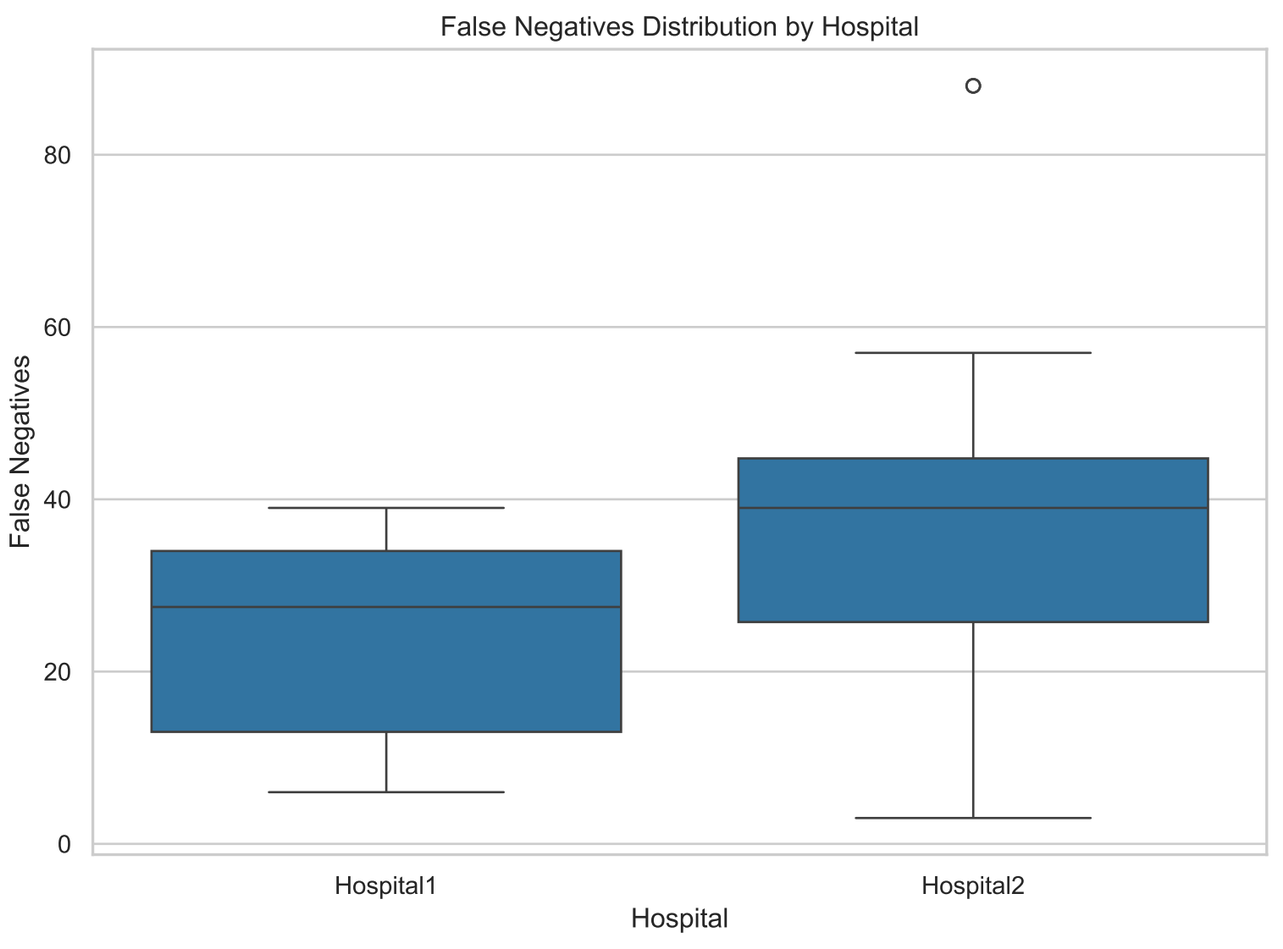

Accuracy Distribution by Hospital
Accuracy
100
80
60
40
20
0
Hospital1
Hospital2
Hospital

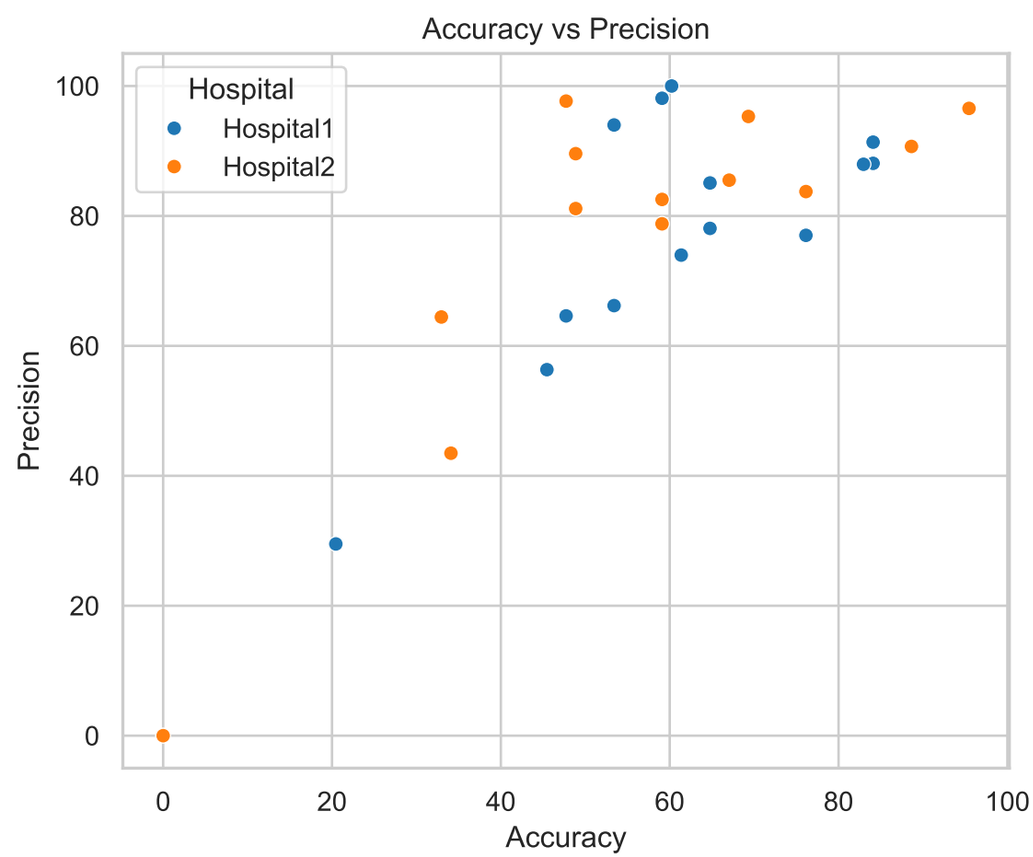

F1 Score Distribution by Hospital
F1score
100
80
60
40
20
0
Hospital1
Hospital2
Hospital

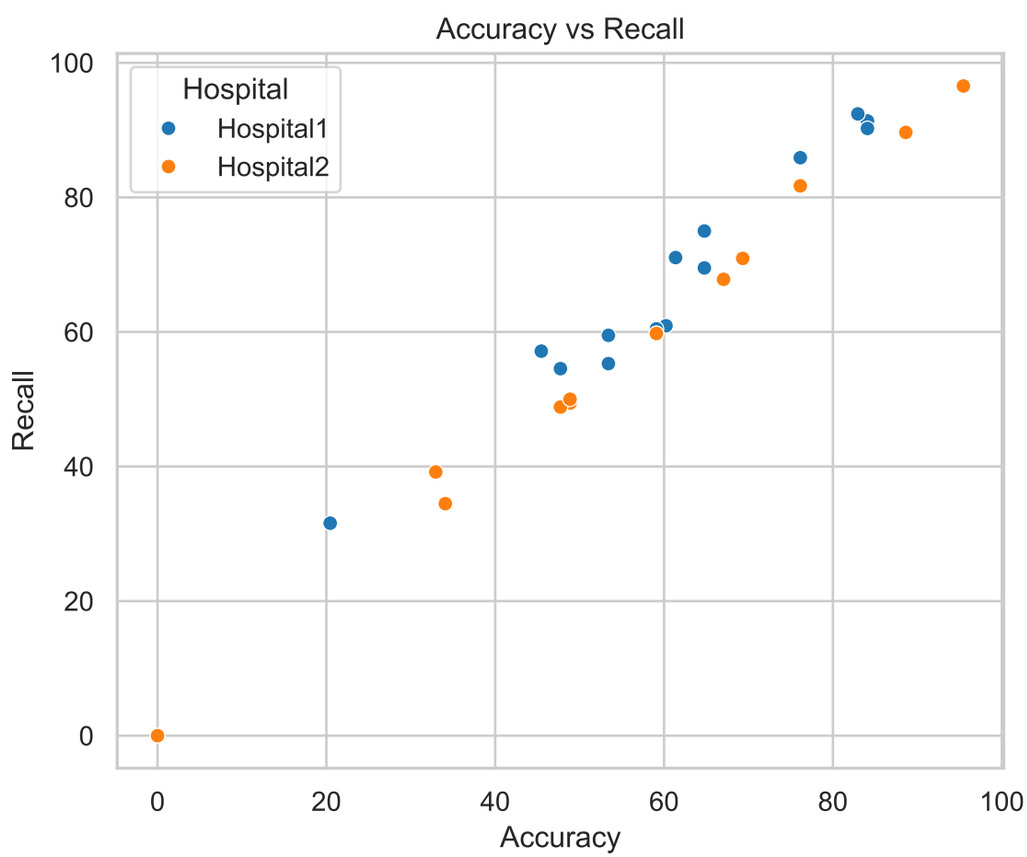

Precision Distribution by Hospital
Precision
100
80
60
40
20
0
Hospital1
Hospital2
Hospital

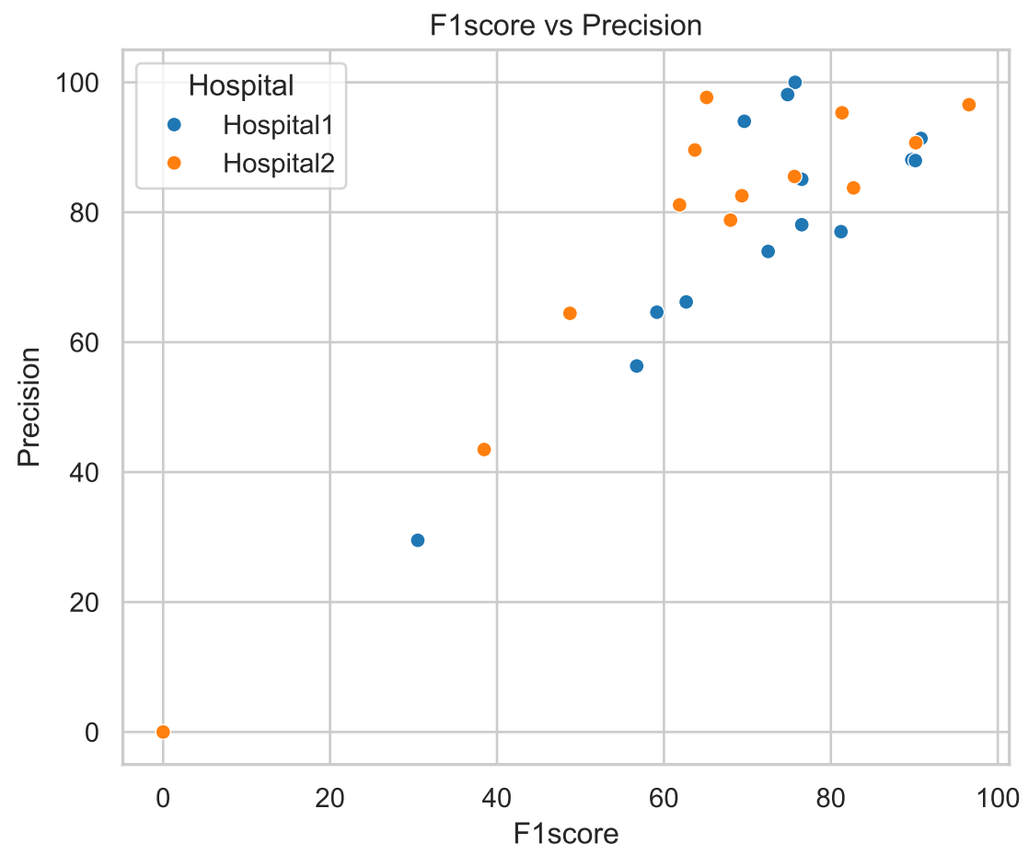

Recall Distribution by Hospital
Recall
100
80
60
40
20
0
Hospital1
Hospital2
Hospital

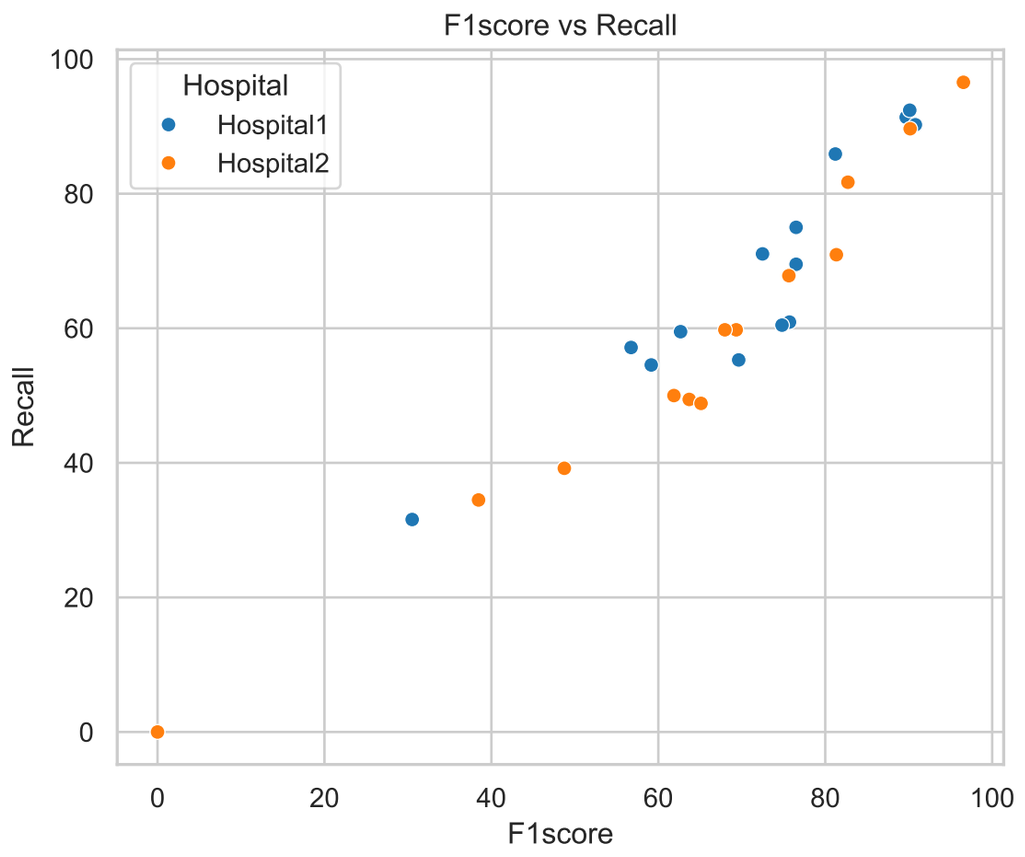

False Positives Distribution by Hospital
False Positives
0
20
40
60
80
Hospital1
Hospital2
Hospital

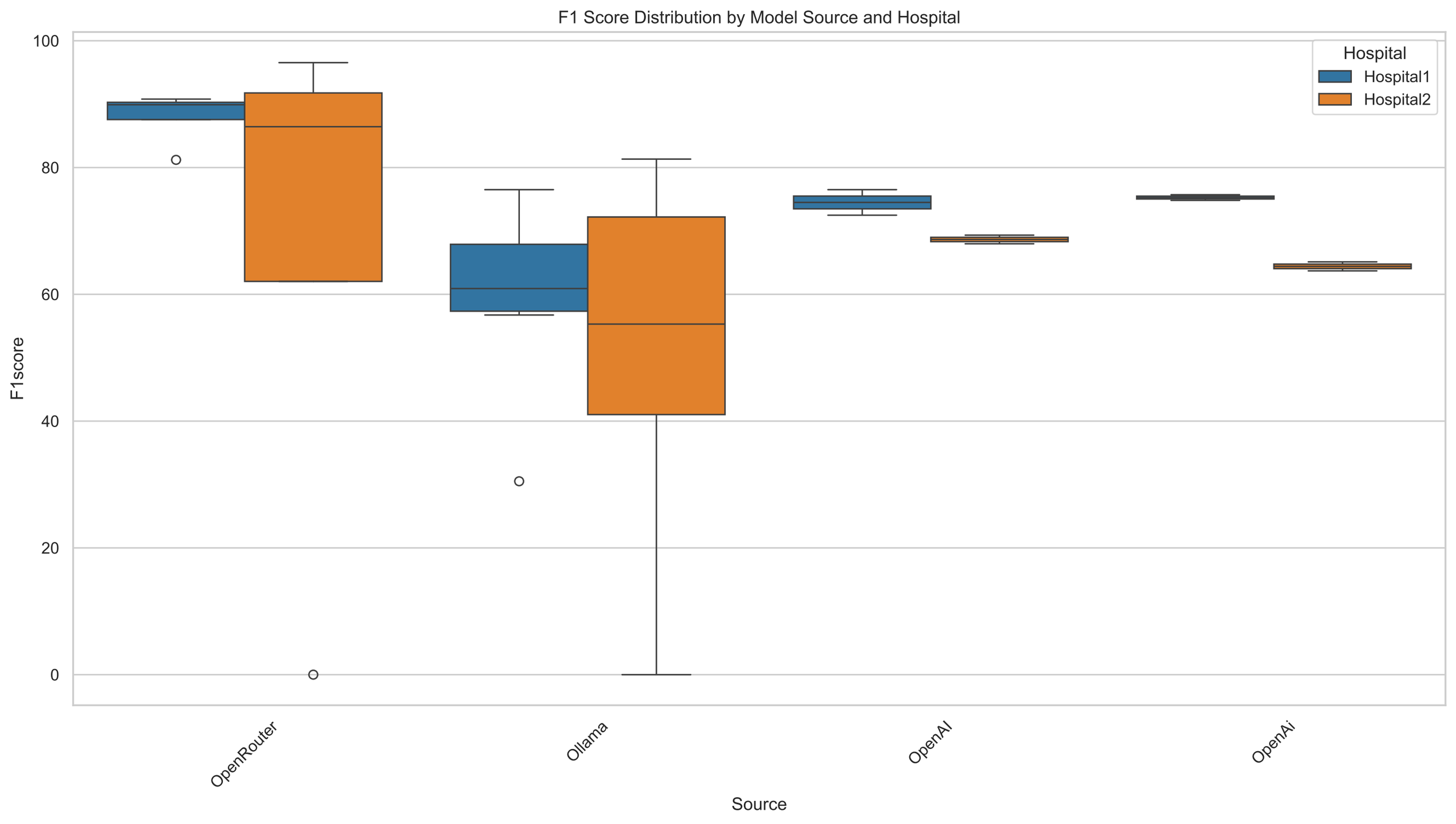

False Negatives Distribution by Hospital
False Negatives
0
20
40
60
80
Hospital1
Hospital2
Hospital

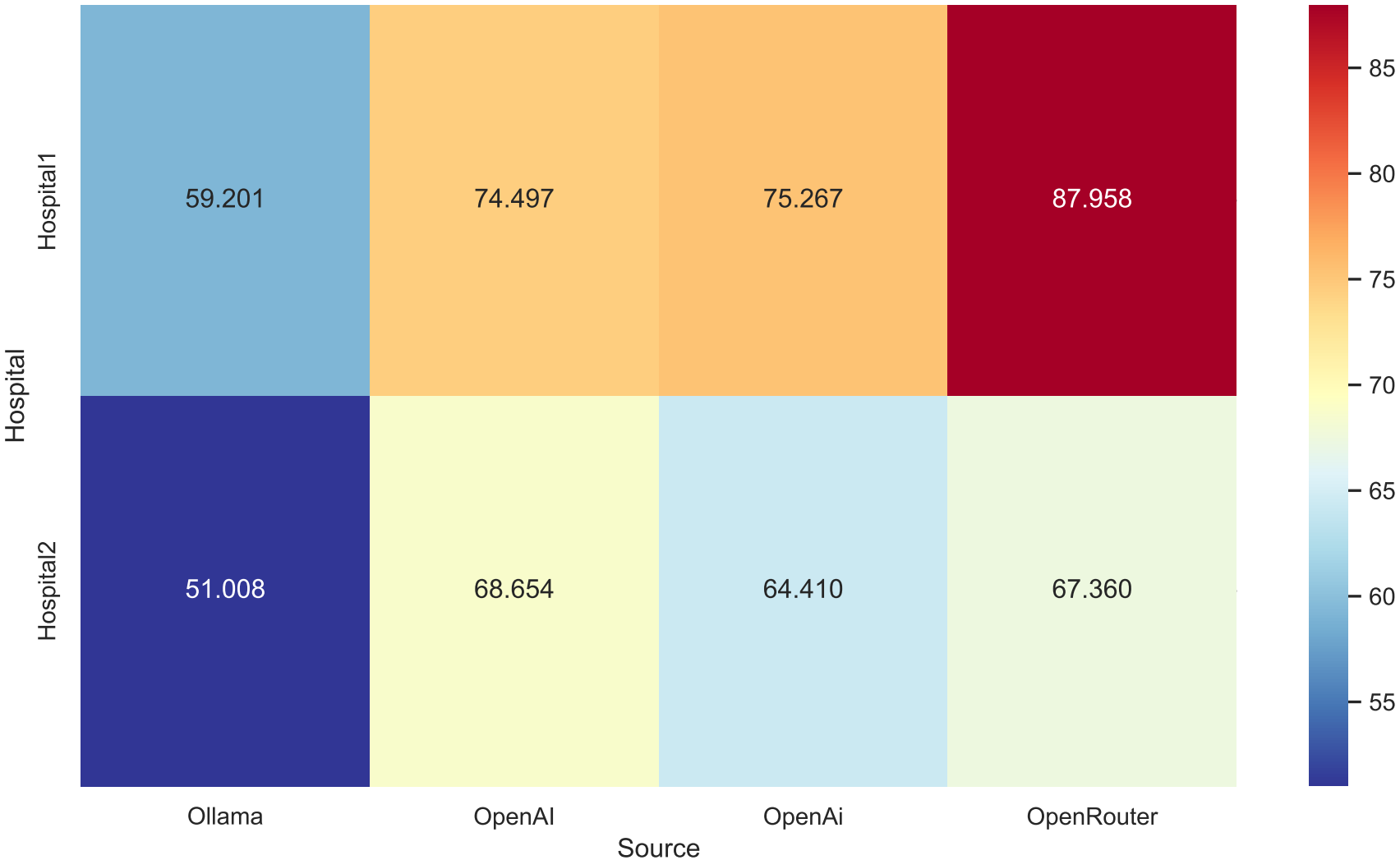

Accuracy vs Precision
Hospital
Hospital1
Hospital2
Precision
Accuracy
0
20
40
60
80
100


Accuracy vs Recall
Hospital
Hospital1
Hospital2
Recall
Accuracy
0
20
40
60
80
100


F1score vs Precision
Hospital
Hospital1
Hospital2
Precision
F1score
0
20
40
60
80
100


F1score vs Recall
Hospital
Hospital1
Hospital2
Recall
F1score
0
20
40
60
80
100


F1 Score Distribution by Model Source and Hospital
Hospital
Hospital1
Hospital2
F1score
100
80
60
40
20
0
OpenRouter
Ollama
OpenAI
OpenAi
Source


Average F1 Score Heatmap: Hospital vs Model Source
Hospital
Hospital1
Hospital2
59.201
74.497
75.267
87.958
51.008
68.654
64.410
67.360
Ollama
OpenAI
OpenAi
OpenRouter
Source
85
80
75
70
65
60
55